Aalto University
School of Science
Master's Programme in Computer, Communication and Information Sciences

Kimmo Kärkkäinen

Predicting demographics and motives of website users

Master's Thesis
Espoo, August 15, 2016

Supervisors: Professor Aristides Gionis, Aalto University
Advisor: Teemu Kinnunen Ph.D. (Tech.)

Aalto University
School of Science
Master's Programme in Computer, Communication and Information Sciences

ABSTRACT OF MASTER'S THESIS

Author:	Kimmo Kärkkäinen		
Title:	Predicting demographics and motives of website users		
Date:	August 15, 2016	**Pages:**	viii + 66
Major:	Computer Science	**Code:**	SCI3042
Supervisors:	Professor Aristides Gionis		
Advisor:	Teemu Kinnunen Ph.D. (Tech.)		

Understanding customers is important. In the case of content providers, this means knowing which groups of people are viewing different types of content, and even more importantly, which groups of people are not yet your active users. However, information about website's visitors is not typically available, as people might visit websites without making an account. Therefore, there is a need for finding out information about the users in other ways.

This thesis studied how accurately demographics (age and gender) and motives can be predicted for website visitors. The predictions were made using the analytics data from a single website. This data included what content the users had viewed, at which times of the day they were active, and what browser and operating system they were using. In addition, this thesis studied how much data is needed to make the predictions, and which features in the analytics data are useful for these predictions.

It turned out to be possible to predict the demographics with good accuracy. Gender was the easiest to predict with 77.8 % accuracy. Age prediction with four age groups had the weighted F1 score of 0.513. Predicting motives was more difficult, and only few of the motives were predicted more accurately than when choosing the most common motives for everyone.

The prediction accuracy depended strongly on how much content the users had viewed. Also, the other features in the analytics data were very weak predictors for the demographics, and therefore they did not increase the prediction accuracy significantly after the user had viewed some content.

Keywords:	demographics, motive, website, machine learning, analytics
Language:	English

Aalto-yliopisto
Perustieteiden korkeakoulu
Tieto-, tietoliikenne- ja informaatiotekniikan maisteriohjelma

DIPLOMITYÖN TIIVISTELMÄ

Tekijä:	Kimmo Kärkkäinen		
Työn nimi:	Verkkosivuston käyttäjien demografioiden ja motiivien ennustaminen		
Päiväys:	15. elokuuta 2016	**Sivumäärä:**	viii + 66
Pääaine:	Tietotekniikka	**Koodi:**	SCI3042
Valvojat:	Professori Aristides Gionis		
Ohjaaja:	TkT Teemu Kinnunen		

Asiakkaiden ymmärtäminen on tärkeää. Sisällöntuottajien kohdalla tämä tarkoittaa ymmärrystä siitä, mitkä ihmisryhmät katsovat erilaisia sisältöjä, ja vieläkin tärkeämmin tietoa siitä, mitä ryhmiä ei ole vielä tavoitettu. Verkkosivun kävijöistä ei kuitenkaan ole usein saatavilla tarkempaa tietoa, koska ihmiset vierailevat sivustoilla ilman rekisteröitymistä. Tämän vuoksi tieto käyttäjistä täytyy selvittää muilla tavoin.

Tämä diplomityö tutki kuinka tarkasti verkkosivun vierailijoiden demografiat (ikä ja sukupuoli) sekä motiivit voidaan ennustaa. Ennusteet tehtiin yksittäisen verkkosivuston analytiikkadataan pohjautuen. Kyseinen analytiikkadata sisälsi tiedon käyttäjien katsomasta sisällöstä, mihin aikoihin käyttäjät olivat aktiivisia ja mitä selainta sekä käyttöjärjestelmää he käyttivät. Lisäksi tämä diplomityö tutki kuinka paljon käyttödataa ennusteiden tekeminen vaatii sekä mitkä piirteet ovat hyödyllisiä ennustamisessa.

Demografioiden ennustaminen osoittautui mahdolliseksi hyvällä tarkkuudella. Sukupuolen ennustaminen oli helpointa, ja sen tarkkuus oli 77,8 %. Iän ennustaminen neljää ikäryhmää käyttäen sai painotetun F1-arvon 0,513. Motiivien ennustaminen oli hankalampaa, ja vain osa motiiveista pystyttiin ennustamaan tarkemmin kuin mitä suosituimman vastauksen valitseminen kaikille olisi antanut.

Ennustuksen tarkkuus riippui vahvasti käyttäjän katsoman sisällön määrästä. Lisäksi analytiikkadatassa vain käyttäjien katsoma sisältö oli merkittävä tekijä, eikä muiden piirteiden lisääminen parantanut tarkkuutta merkittävästi, kun käyttäjä oli katsonut jotakin sisältöä.

Asiasanat:	demografia, motiivi, verkkosivusto, koneoppiminen, analytiikka
Kieli:	Englanti

Acknowledgements

First, I would like to thank my instructor Teemu Kinnunen and professor Aristides Gionis for the feedback and ideas they have given me throughout this project. It was extremely important for the success of this thesis. In addition, I want to thank Eric Malmi for his ideas at the beginning of this project.

I also want to thank Yle, and especially Anne Hyvärilä and Eija Moisala, for letting me write this thesis for them. It has truly been an interesting project and it has allowed me to learn a huge amount of new things.

In addition, I want to thank Futurice for giving me time off from other projects to work on this thesis. Especially, thank you to Anniina Lehtinen, who helped me find this thesis project.

Lastly, special thank you goes to Ela for all the support throughout this lengthy process. It would have been much more difficult without you.

Espoo, August 15, 2016

Kimmo Kärkkäinen

Abbreviations and Acronyms

FN	False negative
FP	False positive
MAE	Mean absolute error
NMF	Non-negative matrix factorization
PCA	Principal component analysis
RMSE	Root mean square error
SVD	Singular value decomposition
SVM	Support vector machine
SVR	Support vector regression
TN	True negative
TP	True positive
TSVD	Truncated singular value decomposition

Notations

x	Scalar value
$\mathbf{x}$	Vector
$\mathbf{X}$	Matrix
$\mathbf{x}^T$	Transpose of vector $\mathbf{x}$
x_i	Value of vector $\mathbf{x}$ at index i
x_{ij}	Value of matrix $\mathbf{X}$ at index (i,j)
$\mathbf{x}_i$	Row i of matrix $\mathbf{X}$
$\|\mathbf{x}\|$	Norm of vector $\mathbf{x}$
x, $\mathbf{x}$, $\mathbf{X}$	Input
y	Output
r	Desired output
$\mathbf{w}$	Weights
b	Bias
C_i	Class i
$p(X)$	Probability of X
$p(X\|Y)$	Conditional probability of X given Y
$log(x)$	Logarithm of x

Contents

Chapter 1

Introduction

1.1 Motivation

It is important for content providers to understand their audience in depth. For example, Netflix collects a substantial amount of data about what their users are viewing each day. With this data, they can predict whether a new show will be a hit or a miss even before it has been launched. They are also able to take it a step further and produce new content based on the known preferences of their user base. A few years ago, they were able to see from their data that their users liked David Fincher, Kevin Spacey, and the British version of House of Cards. Combining these pieces of information gave birth to the American version of House of Cards, which became one of Netflix's most popular series. [Carr, 2013]

In addition to knowing users' preferences, content providers can utilize the information about what the demographics and motives of their users are. With this knowledge, they can understand who the current users of the service are and what types of content different groups of users might prefer. This kind of information also shows which groups have yet to be reached and could be targeted better with new kinds of content or more directed advertisement. This thesis will explore potential methods of predicting more useful information about users based on the information that is typically already available for the content providers.

Demographics are defined as the qualities of a specific group of people, such as age, gender, and income [Merriam-Webster, 2016]. These are qualities that can be measured in a precise way for the majority of people. In this thesis, only age and gender are considered, but the same approaches could be applied to any other demographic data as well. Demographics are the most commonly used tool for customer segmentation, as they are easy to

collect. There is also a wide range of research done on demographic groups (e.g. Li et al. [1999], Bigne et al. [2005]), so the relationship of demographics to other kinds of variables (such as media use) is well known. Tynan and Drayton [1987] note, however, that there is some disagreement on whether demographics are good predictors of behaviour by themselves. They suggest that demographics could be expanded by adding information about the life cycle, such as the number of years married, the age of children, and working status. These features are not examined in this thesis, but the same approaches apply to predicting them.

To understand users even more extensively, it is also important to understand their motives when they use the service. Motives mean the reason why the user came to the service. For example, some user might come looking for content that makes her relax while another user might wish to find current information about what is happening around the world. The same user can be looking for different types of content depending e.g. on the time of the day and the device they are using. This means that the user could be looking for informative content in the morning with their mobile phone while they are on their way to work and entertaining content in the evening using a laptop. Even though these behavioral patterns can be similar for a large portion of the users, there could also be individual differences. Therefore, the knowledge of these can be used to recommend different types of content for users depending on what they are assumed to be looking for. For example, people who typically look for informative content in the evening could get the latest news shows shown to them first when they enter the content provider's website.

Currently, content providers do not have an easy way to find out demographics or motives of their users. The primary way is to ask the users to give them when they are registering to the service. However, in some cases content providers might not want to enforce registration, as it can limit the amount of users. Another way is to send out questionnaires to the users. This, however, requires more time and effort, and usually only a small portion of the users will respond to a questionnaire. The users will also get tired of answering questionnaires if they are sent out too often. Prediction of the needed information could give access to it faster and in a more cost-effective way.

1.2 Yle

Knowledge of demographics and motives are especially important for the Finnish public service broadcasting company, Yle, which runs multiple tele-

vision and radio channels, and which in addition has a website and mobile apps which provide access to a wide variety of news articles, television programs, and radio stations. As Yle is a publicly funded organization, they want to provide interesting content for all the taxpayers. For Yle, understanding their audience's demographics and motives can help them choose what kind of content they need to serve more in order to reach the groups of people who are not yet using their services actively.

Currently, Yle has no accurate information on what types of users are viewing a specific video or article. Only a small portion of the active users have registered to the services and many of the registered users have not given their age and gender. Therefore, Yle needs an alternative way to find out which types of users are viewing their content. Predicting demographics and motives could allow them to understand the behavior of different user groups better. For privacy reasons, Yle does not use the predicted information to profile individual users, but only uses aggregated data to understand the behavior of a larger population.

These needs are not unique to Yle, but they apply to any content provider that does not require users to register to their service before using it. The only requirements are that the content provider knows the wanted information for some of the users, and that they know what content each user has viewed. Even if the wanted information is not currently available for any of the users, it is easy to gather for a part of the users with questionnaires. Therefore, even though this thesis uses Yle's data to explore the different approaches, the same solutions could be applicable anywhere.

1.3 Research goals

Much of the earlier research in this field has focused on predicting user demographics using e.g. the mobile apps they have installed, or the textual content they have produced (see Chapter 2). Prediction of motives has not been studied as extensively. There is also very little existing research on how to make predictions using only the analytics data from a single website.

This thesis aims to extend the current knowledge by finding out whether the predictions can be made using the analytics data from a single website. Compared to the earlier studies which have used all of the users' web browsing data, this imposes a significant limit on the amount of data that is available for a user. This thesis will also find out how accurately the motives can be predicted using the analytics data, which has not been done earlier to the best of our knowledge.

The main research question of this thesis is how accurately the demographics and motives can be predicted based on the analytics data about the user. The analytics data in this case contains the viewed articles and videos, the web browser, the operating system, the device type, and the time of the day. In addition to the main research question, this thesis will find out how much the prediction accuracy depends on the amount of content the user has consumed. This thesis will also find out which features in the typical analytics data are useful for the predictions. This thesis will not go into detail about how to design the necessary questionnaires, as it is a separate field of research in itself.

1.4 Structure of the thesis

This thesis will begin by giving an overview of the earlier research on the topic in Chapter 2. After that, the used methods, such as dimensionality reduction, regression, and classification, are explained on a high level in Chapter 3. Then, Chapter 4 explains how the data was collected and preprocessed. The experiments are described in Chapter 5, and finally, their results are shown in Chapter 6 and they are discussed in Chapter 7.

Chapter 2

Related research

This chapter explores the earlier studies that have been done on predicting demographics and user motives. It also explains how this thesis differs from the earlier studies, and what this thesis aims to add to the existing knowledge on these topics.

2.1 Predicting demographics

Understanding how people's behavior on the internet is affected by their demographics is a commonly researched topic. For example, Bigne et al. [2005] studied how demographics, among other features, can be used to predict mobile purchasing behavior. They discovered that age was a good predictor for behavior while gender was not. Li et al. [1999] made a similar study about how well online buying behavior could be predicted from age, gender, income, and education level, but their results ended up being weak.

Many studies also try to do the opposite, which is predicting demographics from the actions taken by the users. For example, Peersman et al. [2011] used Netlog (Belgian social network) message contents to predict the age of the users. They used support vector machine as the prediction method, and they reached the accuracy of 71.3 % for the age classification task using two age groups (under-16 versus over-16).

Chen et al. [2015] made a study on Twitter users by making predictions of demographics using usernames, self-descriptions, social networks, and profile images in addition to the tweet contents. They used multiple machine learning methods, including support vector machine, decision trees, and logistic regression. For gender prediction, their accuracy was 87.5 % at best, while for age prediction with three age groups they reached the F1 score of 0.595.

Malmi and Weber [2016] researched how to predict age, gender, race, marital status, existence of children, and income level based on what mobile apps the person had installed on their phone. In their study, each of the predicted variables was divided into two categories (e.g. male or female, under 32 years old or over 32 years old), so that they could be predicted with logistic regression. For age prediction they reached the accuracy of 77.1 %, and for gender prediction their accuracy was 82.3 %.

Malmi and Weber also tried reducing dimensionality by i) choosing only the apps that were installed by at least 10 % of the users, ii) aggregating apps based on their categories, and iii) using truncated singular value decomposition (TSVD). None of these resulted with as good accuracy as using the installed apps as features themselves, but TSVD with 500 components reached almost the same accuracy already. They noted, however, that the logistic regression implementation they were using supported sparse matrices, so using TSVD might not provide any additional benefit with regards to space complexity.

Hu et al. [2007] predicted the age group and gender based on visited web pages. They started by giving each web page probability distributions for age groups and genders based on what some of the visitors had self-reported. They then predicted users' demographics based on the demographic distributions of the webpages that they had visited. To make predictions on unseen pages, they also categorize pages based on the text on the page and train the classifier with the categories instead of using the pages directly as features. They then used Bayesian framework to make predictions.

Hu et al. reached the macro F1 score of 0.797 for predicting the gender and 0.603 for predicting the age using five age groups. The most difficult age groups to predict were the youngest and the oldest groups, from which many were classified as belonging to the neighboring group. They also tried reducing dimensionality with latent semantic indexing, which uses singular value decomposition, but this approach did not reach as good results as prediction without it.

As seen, there are numerous studies related to demographics. Primarily, this thesis aims to extend the field of existing knowledge by using analytics data from a single website. Earlier studies have shown how well the predictions can be made if there is data about all the websites that the user visits, but that data is not typically available for the website owners. They can only track what the user does on their website, so it is interesting to find out if that knowledge alone can be enough.

Secondly, this thesis will attempt to find out whether other information available in the analytics data can be used. This includes for example the web browser, operating system, and times of day the user is active. Earlier

studies have primarily focused on the user's actions and content produced by them instead of using these features that are more commonly available.

Thirdly, this thesis will study how the performance of simple learning algorithms compares to a more advanced algorithm. The simple algorithms that are used include linear regression and logistic regression while the more advanced algorithm is support vector machine.

2.2 Predicting motives

Wood et al. [2002] discovered that 54 % of the media usage is habitual, which makes it reasonable to believe that the media usage patterns and motives repeat mostly in similar ways. For example, some person could have a habit of checking the news first thing in the morning with her mobile phone.

Papacharissi and Rubin [2000] and Leung [2006] have studied the motives of internet users. Both of the studies performed a survey asking about internet usage and motives. They then performed a statistical analysis to learn which motives correlated with which types of internet use in general, and they showed that there are some correlations.

Adar et al. [2008] studied how different types of webpages are revisited. They focused on how often different types of content are viewed again (e.g. hourly, daily, or less often). This includes all types of webpages, without being limited to a single website. They showed that different types of websites have different revisitation patterns, and commented that these patterns could correlate with the intents also.

User's context is strongly related to predicting motives also, as habits can depend on the situation the user is in. For example, the user could have a habit of checking specific kinds of content while sitting on the bus. Mayrhofer et al. [2003] tried to find out the user's context based on the data that can be gathered from their mobile phone. This data included for example the time, the noise level, and the number of people using the same wireless connection. Based on these variables, they were able to see when the user's context has changed.

As can be seen from the earlier studies, motives and website revisitation patterns have been studied, but the motives have not been predicted using the web analytics data to the best of our knowledge. This thesis will study whether the prediction is possible by using the content that the user has typically viewed at different times, the content categories, as well as the device types. As the earlier studies have found some correlation between the motives and internet usage, this prediction could be possible.

Chapter 3

Methods

This thesis studies whether demographics and motives can be predicted from web analytics data. This chapter explores the different types of methods that are needed to make these predictions. This includes explaining how features can be engineered, how the dimensionality of the data can be reduced, how the learning is done in the cases of regression and classification, how the parameters are selected, and how the models are compared and selected.

3.1 Feature engineering

Before any machine learning algorithm can be used, the available data needs to be in a form that the algorithms can use. Transforming the data into usable features is called feature engineering [Domingos, 2012]. Domingos argues that feature engineering is the most important part in defining whether the algorithm will succeed in learning or not. He also comments that feature engineering is more complicated than learning the model, as it is highly domain-specific. For example, a spam classifier could use features such as words, word bigrams, and character bigrams [Cormack et al., 2007], whereas music classification into genres could use volume, pitch, and beats per minute as the features [Annesi et al., 2007].

3.2 Information fusion

As data can be gathered from many different sources, there is a need for a method for combining them. Ross and Jain [2003] present three methods for information fusion: i) feature extraction level, ii) score level, and iii) decision level. Feature extraction level means that the feature vectors, which are obtained from different sources, are concatenated into a single vector.

Score level fusion means that a separate model is learned for each feature vector type, and the probability scores given by these models are combined using another model. Decision level fusion is otherwise similar to the score level fusion, except that the models give a prediction instead of a probability score. This thesis uses score level fusion for the classification tasks, as it allows combining very diverse types of data and it takes into account the uncertainty of different models. For regression, decision level fusion is used with another model to combine the decisions.

3.3 Dimensionality reduction

Dimensionality reduction is a class of methods which represent the data using a smaller set of variables. Alpaydin [2010, p. 109] explains that dimensionality reduction lowers the computational cost, and reduction of noise can make the models more robust. Pudil and Novovičová [1998] split the dimensionality reduction methods into ones which are used for optimal data representation and ones that are used for classification. They divide the strategies further into feature selection and feature extraction. Feature selection means that a subset of the original features is used, and feature extraction means using new variables which transform the information contained by the original variables.

For the purposes of this thesis, optimal data representation is not as important as the classification properties. Also, feature extraction is preferred, as it does not lose the information contained by the variables which would be eliminated by feature selection, so it is typically better for discrimination.

Commonly used dimensionality reduction methods include for example principal component analysis (PCA), singular value decomposition (SVD), and non-negative matrix factorization (NMF). Shlens [2014] points out that PCA expects the data to follow a Gaussian distribution along some basis vectors. This assumption does not hold for the data used in this thesis, so PCA is not optimal for this use case.

Xu et al. [2003] compared SVD with NMF, showing that NMF produced easily interpretable clusters automatically while SVD does not. Each axis in the NMF space represents a cluster, and the cluster label of a data point can be determined by the axis along which the value is the highest. This makes analysis of the factorization simple. With SVD, the possibility of negative values makes the analysis more complicated, and a clustering method such as k-means is needed to determine the clusters. Xu et al. also discovered that NMF outperformed SVD-based methods in clustering accuracy. For these

reasons, NMF is used in this thesis, and it is discussed in more detail in the following subsection.

3.3.1 Non-negative matrix factorization

As Xu et al. [2003] present, the purpose of non-negative matrix factorization (NMF) is to find two lower-dimensional matrices $\mathbf{U}$ $(m \times k)$ and $\mathbf{V}$ $(n \times k)$, which could be used to approximate a high-dimensional matrix $\mathbf{X}$ $(m \times n)$ (see Figure 3.1). As an additional constraint, the elements of these matrices must be non-negative. In addition to representing a matrix with two lower-dimensional matrices, NMF is particularly useful for its clustering property. For example, Xu et al. demonstrate that NMF can be used to cluster documents based on their topics.

$$\begin{bmatrix} x_{11} & x_{12} & x_{13} & \dots & x_{1n} \\ x_{21} & x_{22} & x_{23} & \dots & x_{2n} \\ x_{31} & x_{32} & x_{33} & \dots & x_{3n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x_{m1} & x_{m2} & x_{m3} & \dots & x_{mn} \end{bmatrix} \approx \begin{bmatrix} u_{11} & \dots & u_{1k} \\ u_{21} & \dots & u_{2k} \\ u_{31} & \dots & u_{3k} \\ \vdots & \ddots & \vdots \\ u_{m1} & \dots & u_{mk} \end{bmatrix} \begin{bmatrix} v_{11} & v_{12} & v_{13} & \dots & v_{1n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ v_{k1} & v_{k2} & v_{k3} & \dots & v_{kn} \end{bmatrix}$$

Figure 3.1: Matrix $\mathbf{X}$ can be approximated by matrices $\mathbf{U}$ and $\mathbf{V}^T$

As an example, if matrix $\mathbf{X}$ $(m \times n)$ is used to represent which users have viewed which items, m being the number of users and n being the number of items, each row vector shows the items viewed by an individual user, and each column vector shows the users who have viewed a specific item. Matrix value x_{ij} is then 0 or 1, depending on whether user i has viewed item j. After factorizing this matrix to matrices $\mathbf{U}$ and $\mathbf{V}$, the element u_{ik} in $\mathbf{U}$ will represent the degree in which the user i is associated with the cluster k, and the elements v_{jk} in $\mathbf{V}$ will represent the strength of the membership of the item j in the cluster k. With this representation, the matrix $\mathbf{U}$ can be used to see which clusters of items a specific user typically views. This can be faster than using the original matrix directly, and it can also be beneficial for a prediction task, as similar contents are grouped together. Also, matrix $\mathbf{V}$ could be used to verify whether the clustering of items seems sensible.

Xu et al. [2003] show that finding matrices $\mathbf{U}$ and $\mathbf{V}$ can be seen as a constrained optimization problem. The objective is to minimize the distance between the approximation $\mathbf{U}\mathbf{V}^T$ and the original matrix $\mathbf{X}$:

$$J = \frac{1}{2}\|\mathbf{X} - \mathbf{U}\mathbf{V}^T\| \tag{3.1}$$

where $\|.\|$ is the squared sum of the elements in the matrix. The minimization constraints are $u_{ij} \geq 0$ and $v_{xy} \geq 0$, meaning that all the elements in $\mathbf{U}$ and $\mathbf{V}$ should be non-negative. Hoyer [2002] also demonstrates that a sparse solution can be found by including regularization terms into the objective function.

As the problem is Hsieh and Dhillon [2011] show that one way to solve this minimization problem is to use coordinate descent. In this method, $\mathbf{U}$ and $\mathbf{V}$ are updated separately by using the partial derivative to tell in which direction to change the values, while the other matrix stays fixed. Hsieh and Dhillon also demonstrate that this can be further optimized by updating only the most important variables in one matrix before moving to the other one.

3.4 Classification

The objective of classification is to predict which class a data point belongs to. For example, this can mean recognizing which character is in a picture, or whether an email is spam or not. In this thesis, classification is needed to predict gender, age group, and intentions. There are numerous different methods for classification, but this thesis focuses only on two of them: logistic regression and support vector machine. Logistic regression was selected for its simplicity, while support vector machine was chosen as it is a more advanced classification method.

3.4.1 Logistic regression

Logistic regression is a classification method modeling the ratio of class-conditional densities. Modeling class-conditional densities directly would have a significantly larger amount of adjustable parameters, which would cause difficulties in a high-dimensional feature space, as the number of parameters depends quadratically on the number of dimensions. In logistic regression, the parameter count depends on the number of dimensions only linearly. [Bishop, 2006, p. 205]

Figure 3.2 demonstrates how the output values change based on the class densities. In this figure, crosses and dots represent examples from two different classes, and the red line represents the prediction function. As can be seen, the output value changes continuously, and when it reaches the value of

0.5, the probability of the second class becomes higher than the probability of the first class. The class with the highest probability is then chosen as the predicted class.

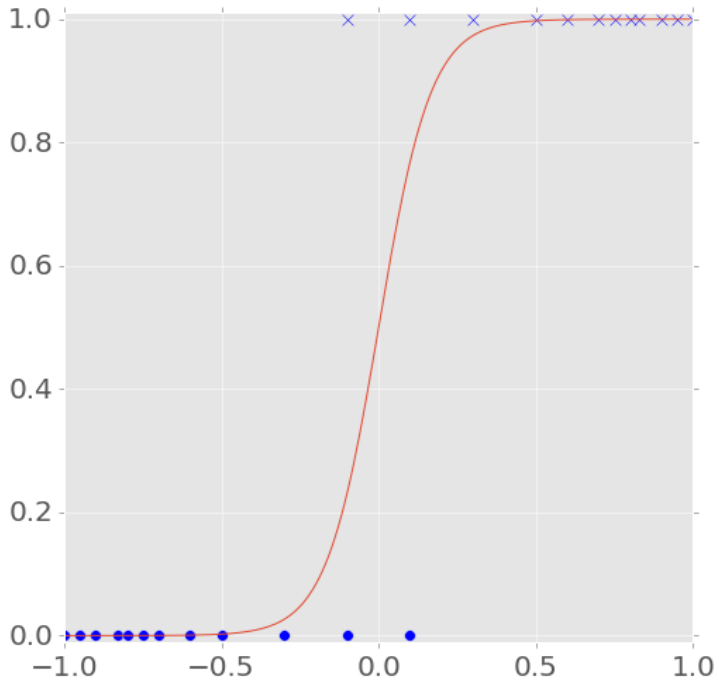


Figure 3.2: Example of logistic regression

Alpaydin [2010, p. 220–221] explains that logistic regression makes the assumption of the log likelihood ratio being linear:

$$log\frac{p(\mathbf{x}|C_1)}{p(\mathbf{x}|C_2)} = \mathbf{w}^T\mathbf{x} + b \tag{3.2}$$

Here, $p(\mathbf{x}|C_1)$ is the class-conditional probability density function of value x given class C_1, $\mathbf{w}$ is the weight vector, and b is the bias value. Using Bayes' rule and rearranging the terms gives the posterior probability:

$$\mathbf{y} = P(C_1|\mathbf{x}) = \frac{1}{1 + e^{-\mathbf{w}^T\mathbf{x}+b}} \tag{3.3}$$

This function can be used to get the predictions once the weights are known. To find out the weights, an error function has to be determined. Alpaydin shows that if the classes are assumed to follow a Bernoulli distribution, the error function E to be minimized is:

$$\begin{aligned} E &= -\log \prod_t (y_t)^{r_t} (1-y_t)^{1-r_t} \\ &= -\sum_t r_t \log y_t + (1-r_t) \log(1-y_t) \end{aligned} \tag{3.4}$$

Here, r_t represents the desired value of example x_t. This minimization problem does not have a closed-form solution, but Alpaydin demonstrates that it can be solved by using gradient descent. In gradient descent, the weights are repeatedly changed in the opposite direction of the gradient until the solution converges.

3.4.2 Support vector machine

Cortes and Vapnik [1995] introduced the currently used version of support vector machine (SVM). The main idea of SVM is to transform the input vectors into a high-dimensional feature space and to form a linear decision surface in this new feature space. It also attempts to maximize the distance between the decision surface and the training vectors. The transformation is done because it is assumed that performing linear discrimination is easier in a high-dimensional space.

Figure 3.3 shows an example of how SVM forms the decision surface for a separable data set. The crosses and dots represent data points belonging to two different classes, and the red line represents the decision surface. Burges [1998] explains that the decision surface depends only on the data points that are nearest to it (shown larger), while the further points can be ignored. The points that define the classification surface are called support vectors.

As shown by Cortes and Vapnik, the goal of SVM is to find a hyperplane which maximizes the distance between the hyperplane and the training vectors. In the separable case, the definition of the separating hyperplane becomes:

$$y_i(\mathbf{w}^T \mathbf{x}_i - b) \geq 1 \tag{3.5}$$

Here, y_i is ± 1 depending on which class the training example $\mathbf{x}_i$ belongs to, $\mathbf{w}$ is the weight vector of the hyperplane, and b is the bias value. The weights are constrained by minimizing:

$$J = \frac{1}{2}(\mathbf{w} \cdot \mathbf{w}) \tag{3.6}$$

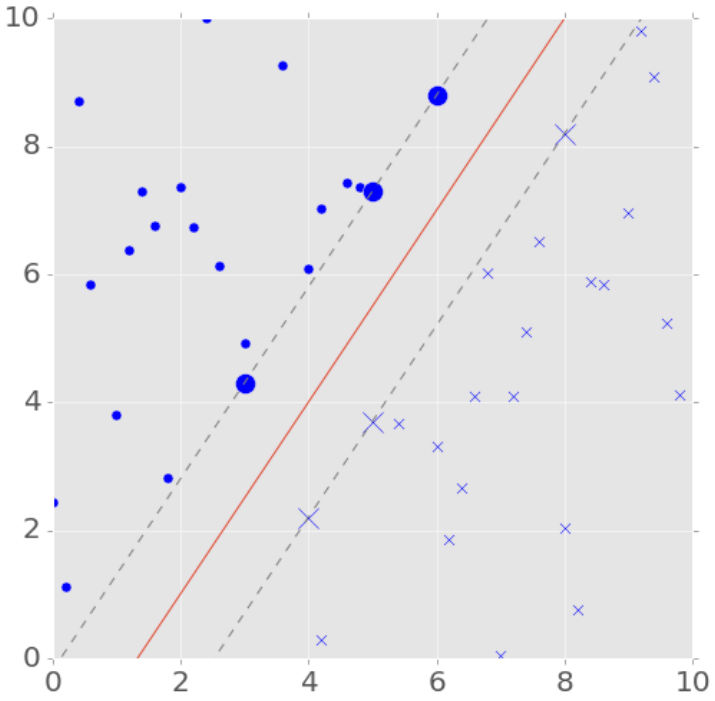


Figure 3.3: Example of support vector machine

Using these formulas, a Lagrangian can be formed:

$$L = \frac{1}{2}(\mathbf{w} \cdot \mathbf{w}) - \sum_{i=1}^{\ell} \alpha_i \Big(y_i(\mathbf{w}^T \mathbf{x}_i - b) - 1 \Big) \tag{3.7}$$

Here, α_i is the Lagrange multiplier. As Cortes and Vapnik show, taking the partial derivatives of the Lagrangian leads to the following definition for the optimal weights:

$$w_o = \sum_{i=0}^{\ell} y_i \alpha_i^o x_i \tag{3.8}$$

Here, α_i^o is non-zero only for the support vectors. This renders it unnecessary to calculate the weights explicitly, as the optimal hyperplane can be determined by using the Lagrange multipliers and support vectors.

However, often the data is not fully separable. Cortes and Vapnik show that SVM can be modified to use a soft-margin optimal hyperplane instead.

This means that errors are allowed, but they have a cost that is added to the optimized function:

$$J = \frac{1}{2}(\mathbf{w} \cdot \mathbf{w}) + C(\sum_{i=1}^{\ell} \xi_i) \tag{3.9}$$

Here, C is a given parameter that determines how much the errors affect the resulting hyperplane. Otherwise, the formulation of the hyperplane is similar to the fully separable case. Burges [1998] explains that these optimization problems can be solved by using any quadratic programming library.

Cortes and Vapnik also demonstrate that as the prediction depends on the inner product between support vectors and feature vectors, the inner product can be replaced by a kernel that satisfies the conditions given by Mercer's theorem (see Mercer [1909]). As Cortes and Vapnik show, kernel functions allow construction of nonlinear decision surfaces. With the usage of kernel functions, there is no need to explicitly map the input vectors into a high-dimensional space, as the same operation can be performed easier with the kernels.

As an example, Cortes and Vapnik show that a radial basis function kernel $K(\mathbf{x}, \mathbf{x}_i)$ can be formed as:

$$K(\mathbf{x}, \mathbf{x}_i) = exp\Big(- \gamma |\mathbf{x} - \mathbf{x}_i|^2 \Big) \tag{3.10}$$

Here, γ defines the radius of the kernel function. With a radial basis function kernel, the proximity to support vectors affects the resulting classification.

3.4.3 Multiclass classification

The classification methods introduced in the previous subsections can only be used for 2-class problems by default. However, there are multiple methods for extending them to work with multiclass problems. Bishop [2006, p. 182–184] shows two methods: one-versus-one and one-versus-rest. In one-versus-one, one classifier is constructed for each pair of classes, and a vote between the classifiers determines the chosen class. In one-versus-rest, one classifier is constructed for each class, and the class with the highest score is selected. However, Bishop also notes that both of these methods suffer from ambiguous areas. For example, two classes might get the same number of votes or the same confidence scores.

This thesis uses the one-versus-rest method, as it requires learning fewer models. For example, in the case of a 4-class problem, one-versus-rest forms 4 classifiers while one-versus-one forms 6 classifiers. This difference can be meaningful when working with large datasets, as learning even one classifier can take a long time. With more classes, the difference becomes even more significant.

3.5 Regression

Regression differs from classification in the sense that the target variable is continuous while classification could only predict prespecified, discrete values [Bishop, 2006, p. 137–138]. In this thesis, regression is used to predict the age of the user. The regression methods used in this thesis are linear regression and support vector regression. Linear regression was selected for its simplicity, while support vector regression was chosen for being a more advanced method.

3.5.1 Linear regression

In linear regression, the objective is to find a function that takes the input vector and produces an output which is as close to the 2true target variable as possible [Bishop, 2006, p. 138–139]. This is demonstrated for the one-dimensional input value in Figure 3.4, where the horizontal axis represents the input values, the vertical axis represents the target values, and the crosses represent measured data points. Linear regression learns a function (shown as the red line), which tries to represent the true function as accurately as possible.

Bishop [2006, p. 138–139] shows that in the simplest case, linear regression can be done using the following function $g(\mathbf{x})$ for prediction:

$$g(\mathbf{x}) = \mathbf{w}^T\mathbf{x} + b \tag{3.11}$$

The prediction is made by taking a dot product between the input vector ($\mathbf{x}$) and weights ($\mathbf{w}$), and by adding a bias value (b). The optimal weights are unknown, so they must be learned before making predictions. Because each input variable has a corresponding weight that it is multiplied by, it is easy to see from the weights how a change in some input variables affects the prediction.

To find out what the weights should be, an error function has to be selected for minimization. Bishop [2006, p. 5–6] explains that a typical choice

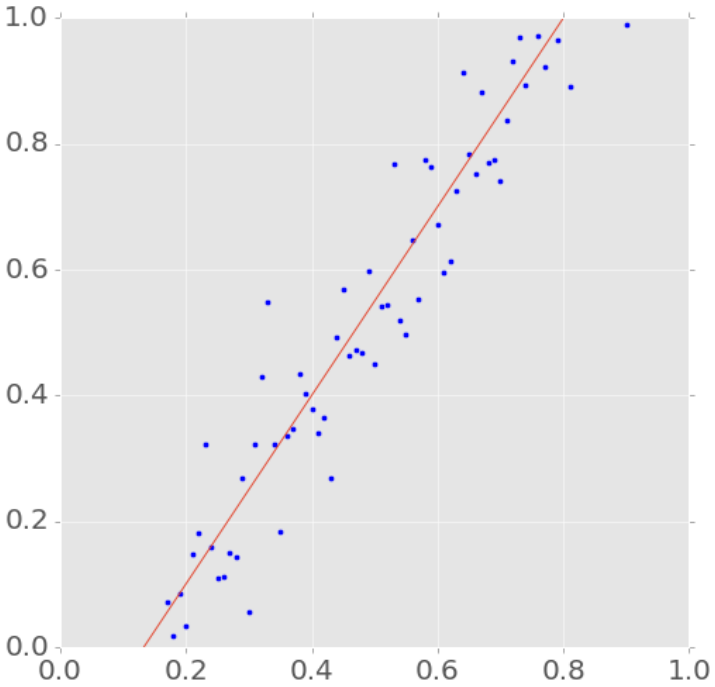


Figure 3.4: Example of linear regression

for an error function is the sum of the squares of the errors, as it provides non-negative values, and as it goes to zero only when the predictions are exactly the same as the real target values. Mathematically, this function can be expressed as:

$$E = \frac{1}{2}\sum_{n=1}^{N}\Big(r_n - g(\mathbf{x}_n)\Big)^2 \tag{3.12}$$

In this equation, N is the number of examples in the training set, r_n is the true output value, and $g(\mathbf{x}_n)$ is the prediction for input $\mathbf{x}_n$. When this is differentiated with regards to the weight vector to minimize the error, the weight vector becomes:

$$\mathbf{w} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{r} \tag{3.13}$$

Here, $\mathbf{X}$ is a matrix where each row is an input vector, and t is a vector containing the corresponding true output values. $(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$ is called Moore-Penrose pseudo-inverse, and it is a generalization of the matrix inverse to non-square matrices. [Bishop, 2006, p. 141–142]

Bishop [2006, p. 143–146] notes that if the dataset has co-linearities, meaning that some of the variables are not independent, there is a risk that the weights can become too large. This is typically not good for generalizability, but it can be avoided with two methods: by adding a regularization term to the objective function, or by removing the dependencies by using a dimensionality reduction method. Adding a regularization term penalizes complex models, but as Alpaydin [2010, p. 80] adds, it can also increase the risk of choosing a too simple model. Therefore, the weight of the regularization term should be optimized using cross-validation.

Bishop shows that the regularization term commonly takes the following general form:

$$\lambda \sum_{j=1}^{M} |w_j|^q \tag{3.14}$$

Here, λ determines how strong the regularization is, M is the number of weights, and q is the chosen regularization type. When $q = 1$, the solution is likely to become sparse, meaning that many of the weights will become zero. With $q = 2$, the solution will have more non-zero weights, but these weights are smaller than with $q = 1$. This thesis uses L2 regularization ($q = 2$), as L1 regularization ($q = 1$) would result in ignoring some of the input values.

3.5.2 Support vector regression

Support vector regression (SVR) is a method using the concept of support vectors (see Section 3.4.2) to approximate continuous function values. It was first introduced by Vapnik [1995], and the main idea of it is to transform the feature vectors into a high-dimensional space where the regression problem becomes linear.

The simplest version of SVR is the ϵ-SVR, which is demonstrated by Smola and Schölkopf [2004]. The purpose of it is to find a function, which is as flat as possible while all of the training examples are at most ϵ distance away from the value given by the function. Flatness means that the weights ($\mathbf{w}$) of the regression function are as small as possible. This regression function is shown in Figure 3.5, where the dots are known data points, the red line represents the prediction function, and the dashed lines show how far the data points are allowed to be.

Smola and Schölkopf [2004] define the regression function as:

$$f(\mathbf{x}) = \mathbf{w}^T\mathbf{x} + b \tag{3.15}$$

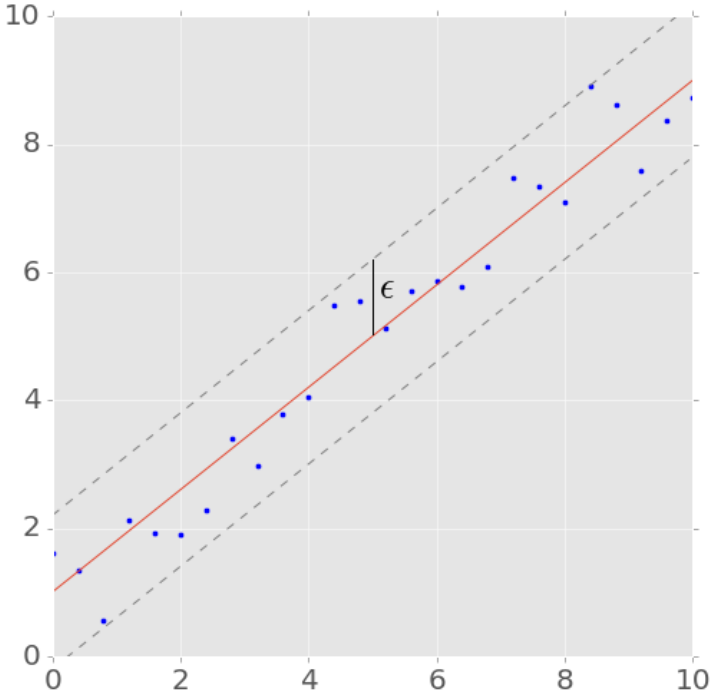


Figure 3.5: Example of ϵ-SVR

Here, $\mathbf{x}$ is the input vector, $\mathbf{w}$ is the weight vector, and b is the bias. To find an optimal solution, the function to be minimized is:

$$J = \frac{1}{2}|\mathbf{w}|^2 \tag{3.16}$$

With the constraints:

$$\begin{aligned} y_i - \mathbf{w}^T\mathbf{x} - b &\leq \epsilon \\ \mathbf{w}^T\mathbf{x} + b - y_i &\leq \epsilon \end{aligned} \tag{3.17}$$

As noted by Smola and Schölkopf, in practice it might not be feasible to have all of the training examples within an ϵ distance from the regression function. We might, for example, want to allow some larger deviations to have better generalizability. To allow larger deviations, the optimization problem can be altered to find a balance between the amount of errors larger than ϵ and the flatness of the regression function. This balance is controlled by the parameter C, which can be seen in the mathematical representation later. Figure 3.6 demonstrates a larger deviation from the regression line. In

the mathematical representation, the deviations are represented differently depending on which side of the regression function the outlier is. On one side, the deviation is represented with ξ_i and on the other side it is represented by ξ_i^*.

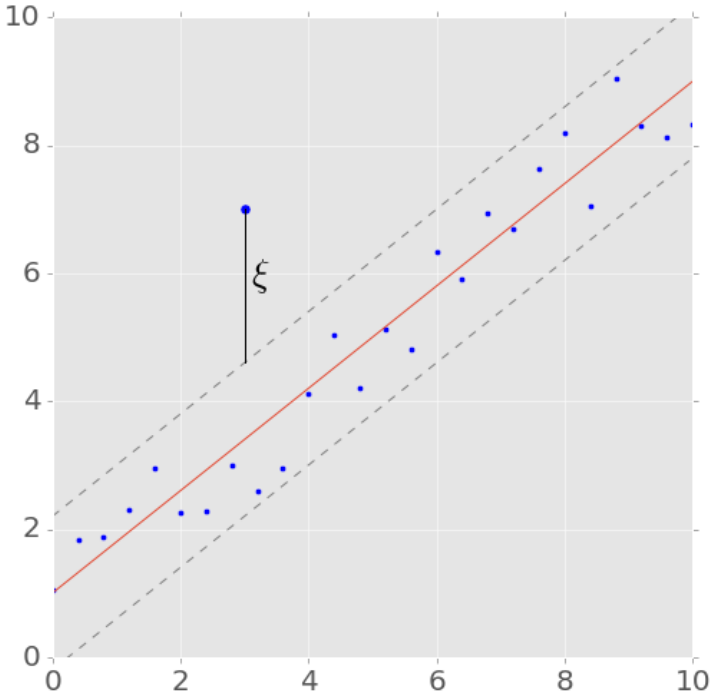


Figure 3.6: Example of SVR with a soft margin

Smola and Schölkopf show that the function to be optimized then becomes:

$$J = \frac{1}{2}|\mathbf{w}|^2 + C\sum_{i=1}^{\ell}(\xi_i + \xi_i^*) \tag{3.18}$$

With the constraints:

$$\begin{aligned} y_i - \mathbf{w}^T\mathbf{x} - b &\leq \epsilon + \xi_i \\ \mathbf{w}^T\mathbf{x} + b - y_i &\leq \epsilon + \xi_i^* \\ \xi_i, \xi_i^* &\geq 0 \end{aligned} \tag{3.19}$$

Smola and Schölkopf continue that the objective function can then be formed as a Lagrangian function where α_i and α_i^* are the Lagrangian multipliers:

$$
\begin{aligned}
L = & \frac{1}{2}|\mathbf{w}|^2 + C\sum_{i=1}^{\ell}(\xi_i + \xi_i^*) - \sum_{i=1}^{\ell}\alpha_i(\epsilon + \xi_i - y_i + \mathbf{w}^T x_i + b) \\
& - \sum_{i=1}^{\ell}\alpha_i^*(\epsilon + \xi_i^* + y_i - \mathbf{w}^T x_i - b) - \sum_{i=1}^{\ell}(\eta_i\xi_i + \eta_i^*\xi_i^*)
\end{aligned} \tag{3.20}
$$

To minimize it, partial derivatives are taken, and it can be seen that the weights can be represented by the training examples and Lagrange multipliers instead:

$$
w = \sum_{i=1}^{\ell}(\alpha_i - \alpha_i^*)x_i \tag{3.21}
$$

Because of this, there is no need to compute the weights, but the whole algorithm can be performed using the training examples instead. As with SVM, the dot products can be replaced with kernel functions, which enables operating in high-dimensional spaces without actually performing the transformations directly. This makes the regression non-linear. Also, the α_i and α_i^* become zero for training examples that are within the ϵ margin, so SVR will end up with a sparse solution using only the training examples that were further away. Solving the optimization problem is also similar to SVM, as it can be done using existing quadratic programming libraries. [Smola and Schölkopf, 2004]

3.6 Parameter selection

Typically, each machine learning method has some parameters that can be tuned to get better results. Parameters could be selected by making assumptions about which ones should work in a specific case, but this is unlikely to produce the most optimal parameters. Slightly better method would be to adjust one parameter to find the optimal value for it and then adjust the next one. This makes the assumption that the optimal parameter values do not depend on the other parameter values, which is not true usually. Grid search solves this problem by trying all the possible combinations to find out the optimal combination of parameters. [Alpaydin, 2010, 480-481] This thesis uses grid search to select the parameters.

3.7 Model selection

To know which model to choose, the performance of different models needs to be evaluated and compared. This means that there needs to be some metric that is used to decide which model has the best performance. Alpaydin [2010, p. 40] also notes that the performance on the training data does not indicate how well the model will perform on new data, so there is also a need for a process to estimate this performance. This subsection goes through some of the commonly used metrics, as well as demonstrates how the performance of a method can be estimated by using a method called cross-validation.

3.7.1 Evaluation metrics

As shown by van Rijsbergen [1979], in a two-class classification problem four values can be counted: true positive (TP), true negative (TN), false positive (FP), and false negative (FN). These are demonstrated in Figure 3.7, where the green area represents the positives examples, red area represents the negative examples, and the lighter area represents the samples that were predicted to be positive. True positive means how many of the data points belonging to the positive class were classified correctly as positive. True negative correspondingly means how many of the data points belonging to the negative class were classified correctly as negative. False positive and false negative mean the numbers of data points that were incorrectly classified into the opposing classes.

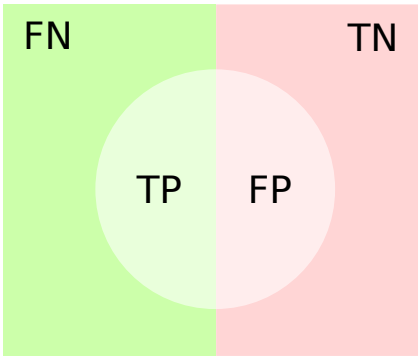


Figure 3.7: Different types of prediction correctness and incorrectness

Van Rijsbergen shows that these values can be summarized in multiple ways based on what is needed. A simple way to do this is to use accuracy A, which tells what portion of the items were correctly classified:

$$A = \frac{TP + TN}{TP + TN + FP + FN} \tag{3.22}$$

Precision P tells what portion of the positive predictions were actually positive:

$$P = \frac{TP}{TP + FP} \tag{3.23}$$

And recall R tells the portion of the positive items that were correctly classified as such:

$$R = \frac{TP}{TP + FN} \tag{3.24}$$

As both precision and recall can be important, van Rijsbergen combined them into a single measure, F-measure, which is the harmonic mean of precision and recall:

$$F = \frac{(1 + \beta^2)(P \cdot R)}{\beta^2 P + R} \tag{3.25}$$

Van Rijsbergen shows that β can be used to adjust the weight of precision and recall to suit the needs better. In the typical case where both should be treated equally, β is set to 1. If a high recall is more important than high precision, β can be higher (e.g. 2), and if precision is more important, β can be lower (e.g. 0.5).

To use these metrics with multiclass classification, these metrics can be calculated separately for each class, and a weighted average can be taken of the results. Using a weighted average means that the more common classes gain more importance than the rare classes.

In regression, different types of metrics are needed. Chai and Draxler [2014] show two commonly used metrics for regression problems: mean absolute error (MAE) and root mean square error (RMSE). Mean absolute error is defined as:

$$MAE = \frac{1}{N} \sum_{i=1}^{N} |e_i| \tag{3.26}$$

Here, N is the number of examples and e_i is the error of the example x_i. And root mean square error is defined as:

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N} e_i^2} \tag{3.27}$$

Chai and Draxler comment that RMSE is the more appropriate metric when the error distribution is assumed to be normally distributed, as the larger errors get more weight. As this assumption is expected to hold with the predictions made in this thesis, RMSE is used as the evaluation metric for regression problems.

3.7.2 Cross-validation

To estimate the prediction quality on new data, Alpaydin [2010, p. 40] proposes that cross-validation is used. Cross-validation means that the dataset is divided into a training set and a validation set. Then the model is trained using the training set and the performance is tested using the validation set. To evaluate the performance more accurately, this needs to be done multiple times with different training and validation sets. Because there typically is not enough data to split it into multiple training and validation sets, a method called K-fold cross-validation is used.

In K-fold cross-validation, the data set is split into K folds, which is demonstrated in Table 3.1. The model is learned using all except one of these folds, and the performance is then evaluated using the fold that was left out. This is repeated K times, so that each fold is used once as the validation set. Choosing a large K increases the training set size, but decreases the validation set size. Also, the model needs to be trained K times, so a large value will mean slower performance. If there is very little data, choosing a large K is required, as the training set would be too small otherwise. An extreme choice is to make K the same as the number of instances in the data set, so that in each round only one example is used in the validation set. This special case is called leave-one-out. [Alpaydin, 2010, p. 486–488]

As the model is selected based on the performance on the validation data, the hyperparameters are selected based on that dataset. Because of this, the results with that dataset are likely to be overly optimistic. To avoid this problem, it is necessary to have one more data set that is not touched during the model selection part, and this dataset is called the test set. When the performance of a model is reported, this is the dataset that is used for evaluating the performance. [Alpaydin, 2010, p. 40]

Table 3.1: Example of 5-fold cross-validation

Round \ Fold	**20 %**	**20 %**	**20 %**	**20 %**	**20 %**
1	Validation	Training	Training	Training	Training
2	Training	Validation	Training	Training	Training
3	Training	Training	Validation	Training	Training
4	Training	Training	Training	Validation	Training
5	Training	Training	Training	Training	Validation

Chapter 4

Data collection and processing

The purpose of this thesis is to predict the demographics and intentions of website's users. Before predictions can be made, the data needs to be collected and preprocessed. This chapter shows how the analytics data and questionnaire data were collected and processed. In addition to describing the processes, the data is also analyzed and relevant statistics are shown.

4.1 System structure

To understand how the experiments in this thesis were performed, it is important to first understand how the complete data collection and processing pipeline works, starting from the website and ending in the predictive model. This pipeline is demonstrated in Figure 4.1.

The pipeline starts with the website, which has an analytics system. The analytics system collects events when users perform specific actions, and these events are stored into a database. These actions included for example user viewing a page, scrolling the page to see a specific element, starting to play a video, and so forth.

Most of the data in the database was not necessary for prediction purposes, so the data was first filtered to contain only the useful parts. This made it possible to process the data much more efficiently. The data was also enriched to include the used browser and operating system based on the user agent data sent by the browser.

Then, the data was transformed into feature vectors, so that the prediction algorithms could use it. Each user had one feature vector, where the contents depended on the experiment that was being performed. For example, a feature vector could show which videos and articles a specific user had viewed, or which devices the user had used to access the service. The exact

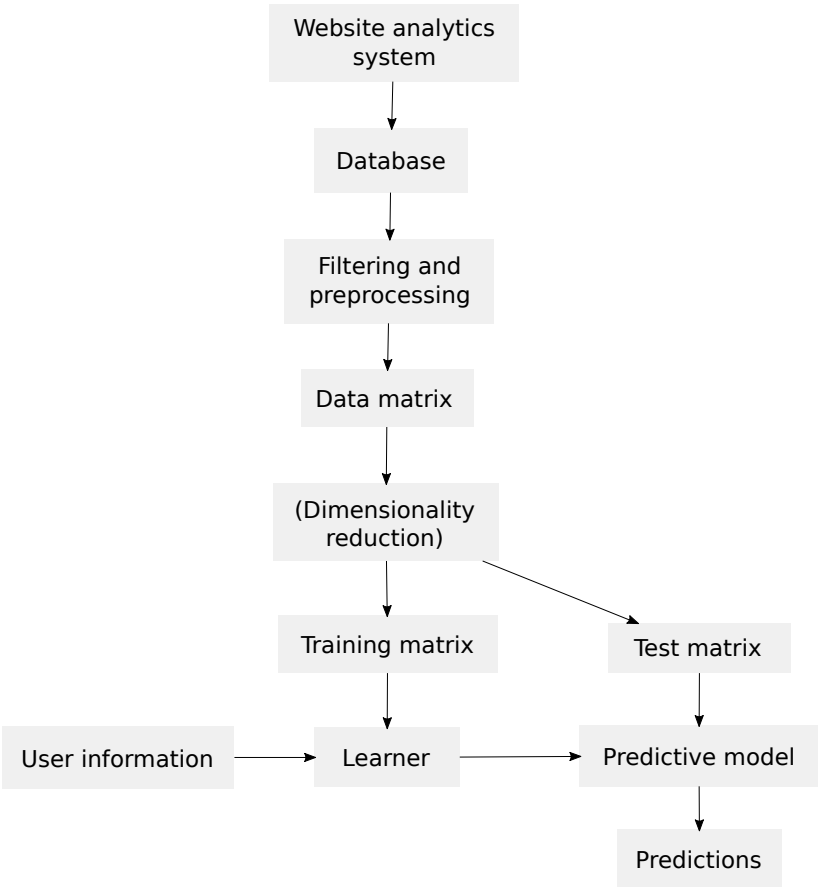


Figure 4.1: Description of the prediction system structure

contents of the vectors are described in more detail later. These vectors were combined into one data matrix.

Next, depending on the experiment, dimensionality reduction was performed on the data. The assumption was that the models could perform better with a lower-dimensional dataset. Dimensionality reduction was not done in all of the experiments, so the usage of it is stated separately in each such case.

After processing the data, it was split into parts: training data and test data. The training data consisted of randomly picked 75 % of the users, and the rest of the users were used in the test data. This split was done to get accurate measurements of the prediction performance on unseen data.

Finally, this matrix was given to the learning algorithm with the correct answers (e.g. users' ages or genders) for the corresponding users, and the model was learned using this data. The learned model was then used to predict demographics or motives for a group of users that were in the test data. These predictions were used to report the performance of the model.

4.2 Analytics data

The media usage data was collected by the analytics system on Yle's websites and mobile applications. The data was not specifically purposed for this study, but it represented the type of data that is typically collected by websites. The data was originally collected into a database, and for the purposes of this study the data was extracted from there for the time period between 26.2.2016–18.4.2016. The extracted data included only events about article views and video plays. In addition, metadata about the articles and videos, such as the subjects given to them, was used.

This thesis focused only on the registered users that had been active during the observed time period. This meant viewing at least five articles or videos. After filtering less active users out, there were 35335 users left. The videos and articles were also filtered based on the view count, so that all items with less than ten views by identified users were left out. After filtering the less popular contents, there were 10890 videos and 9566 articles left.

Because of the limitations of the analytics system, many of the article view events did not include the user's identifier. To solve this problem, all the cookie and user identifier pairs were collected from the dataset, and if an event was missing the identifier, it was inserted based on the cookie if possible. It was then assumed that all the events with the same cookie had come from the same user. Even after adding the known user identifiers, the

video data had three times the amount of identified users compared to the article data, which could cause the prediction accuracy to be lower when using the article data.

Figure 4.2 displays the age distribution of the registered users. The dataset had a relatively large amount of users from each age group, but the ages between 50–70 were clearly overrepresented. The amount of men and women was almost equal, with 51.7 % men and 48.3 % women. The gender and birth year were self-reported at registration time, but they were not required fields, so the data was not available for everyone. Because the demographic data was self-reported with no verification, there was no certainty about the correctness of it.

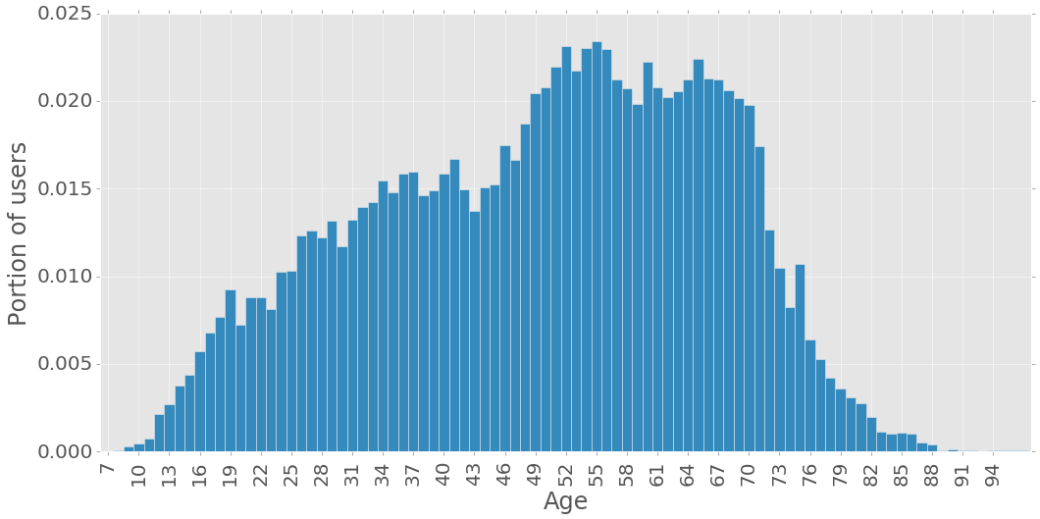


Figure 4.2: Age distribution of the active registered users

The video and article data were handled as bag-of-contents, meaning that each user's feature vector included either one or zero for each video and article depending on whether the user had viewed the video or article. Article and video categories were also treated in the same way. Each user's feature vector was also scaled to have a unit norm.

The operating system, the web browser, and the times of use were represented as vectors where each element showed how large portion of the user's events had happened with that operating system or browser, or how many of the events had happened in a specific time block. The time blocks were 4-hour blocks (02–06, 06–10, 10–14, 14–18, 18–22, 22–02) with weekdays and weekends separated. This meant having 12 time blocks in total. The assumption here was that the majority of the users work on weekdays, so they might not use the service for entertainment during the days. On the week-

ends they could behave differently, as they are not at work. Obviously this approach does not work for everyone, but it was assumed that it would work for a large portion of the population.

4.3 Analysis of the analytics data

To see whether operating system, browser, or times of activity could be useful for prediction, the demographics were plotted for them.

First, the operating systems used by the users were analyzed. Figure 4.3 shows the age distributions and Figure 4.4 shows the gender distributions for it. Clearly, OS X is used more by the younger population while Windows Vista is used more by the older population. Also, Linux is strongly male-dominated, while Windows 8.1 is female-dominated. However, age and gender distributions of the most commonly used browsers and operating systems do not differ from the whole populations distributions significantly. Because of this, they are not likely to be useful for the prediction. The less commonly used browsers and operating systems can be informative, but they are useful only for a small portion of the users. Because of the small differences, it is still worth experimenting on.

Next, the web browsers used by the users were analyzed. The age distributions can be seen in Figure 4.5 and gender distributions can be seen in Figure 4.6. It can be seen, for example, that the users of Internet Explorer or Edge are mostly older people. Also, Edge users are more commonly male than female, similarly to Android mobile application users. Again, the differences are not large, but the minor differences could make using these useful.

Finally, the times of the day when the user was active were analyzed. When looking at age group distributions in different time blocks in Figure 4.7, it can be seen that the behavior is very similar between different age groups. The only noticeable difference between the age groups is that older people tend to start using the service before 6 in the morning more commonly than the younger people. However, this is also relatively more quiet time in the services, so this information is useful only with a small portion of the population. Figure 4.8 shows that the usage patterns are very close to each others between genders. Based on this, the times of activity should not be very useful for predictions, but it was still experimented to verify this assumption.

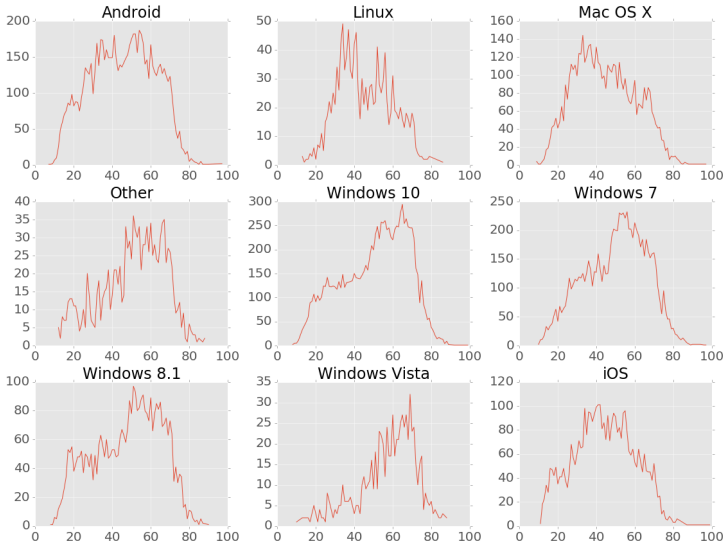


Figure 4.3: Age distributions for different operating systems. Horizontal axis represents the age and vertical axis represents the amount of users.

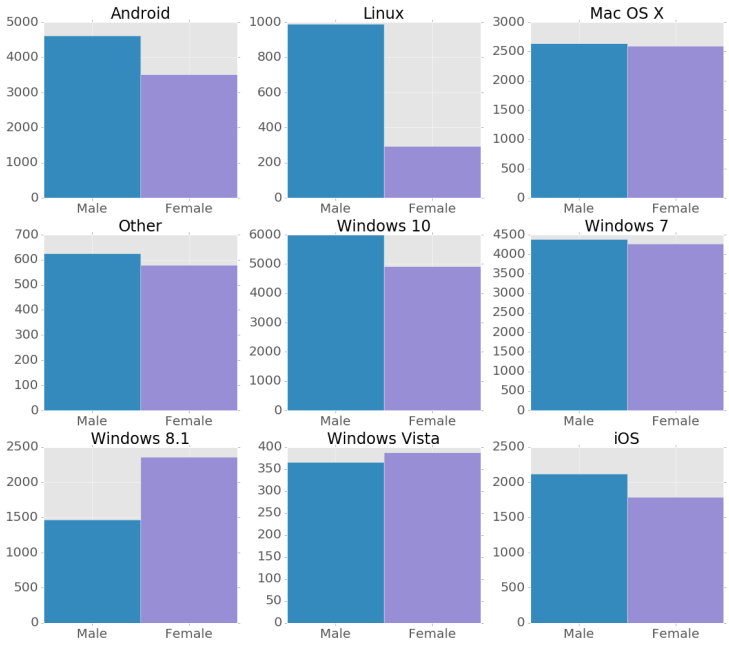


Figure 4.4: Gender distributions for different operating systems. Left bar represents the male users and right bar represents the female users.

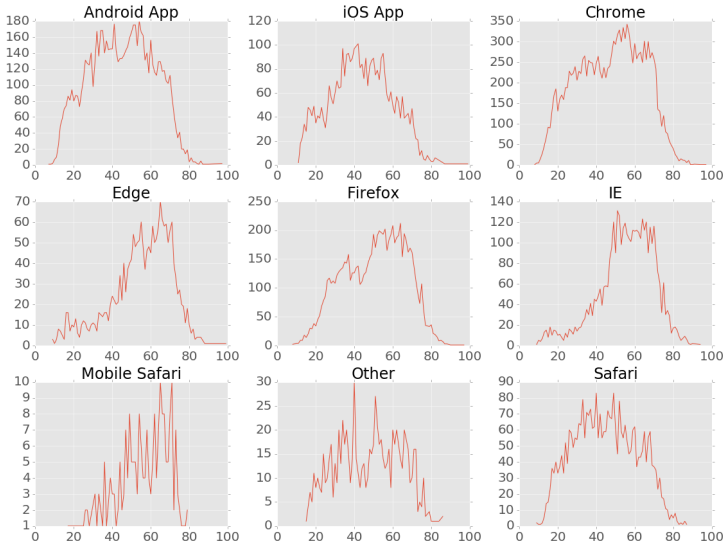


Figure 4.5: Age distributions for different browsers. Horizontal axis represents the age and vertical axis represents the amount of users.

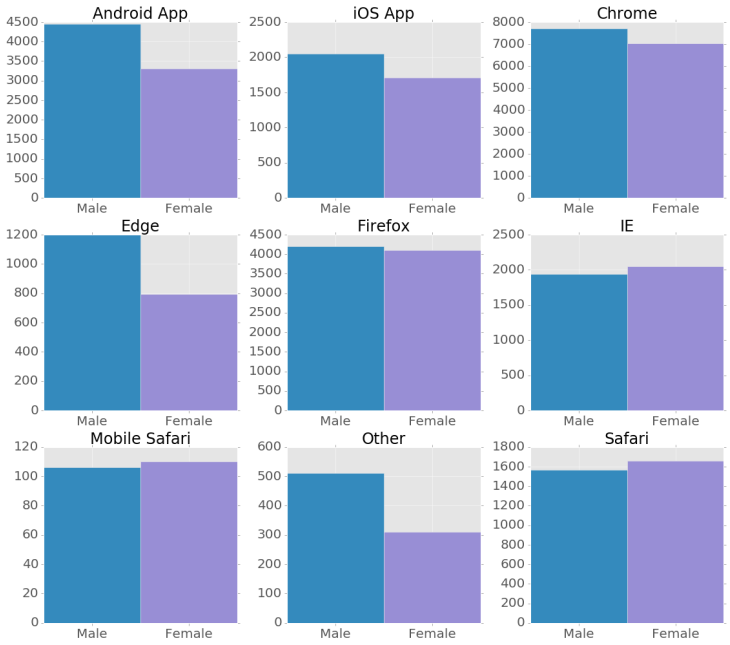


Figure 4.6: Gender distributions for different browsers. Left bar represents the male users and right bar represents the female users.

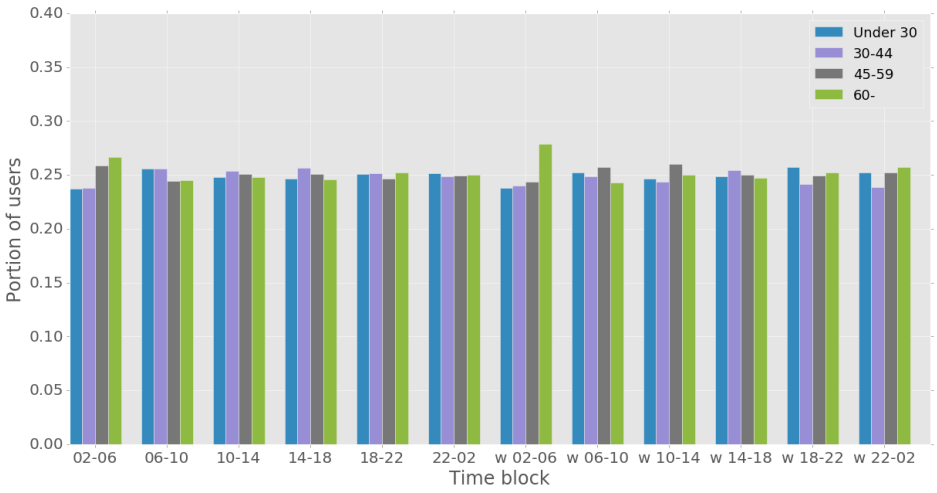


Figure 4.7: Age group distributions at different times of the day. Left half of the figure contains weekdays and right half contains weekends.

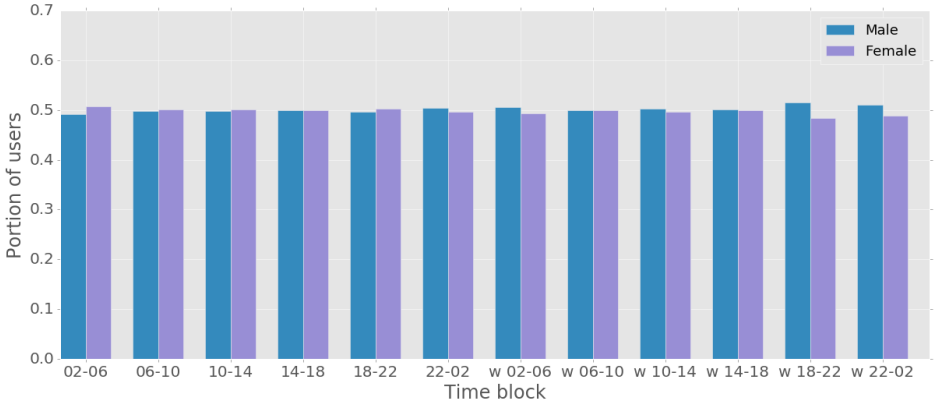


Figure 4.8: Gender distributions at different times of the day. Left half of the figure contains weekdays and right half contains weekends.

4.4 Questionnaire

One of the goals in this thesis was to predict what the motives of the users are when they use the service at different times of the day. To be able to predict this, some training data was needed. To gather it, a questionnaire was sent to a group of active users to find out what their typical motives on weekdays were. The questionnaire consisted of ten proposed motives, and the users could choose any number of them. These motives were asked for four different time blocks: morning (6–9), afternoon (12–15), early evening (18–21), and late evening (21–24). The given options were the following:

- I want to get current information about what is happening in the world
- I want to get current news within my proximity
- I want to deepen my knowledge about interesting things
- I want to enjoy, relax
- I want to become happy, laugh
- I want content that raises emotions or thoughts
- I want content that I can share or forward to others
- I want content that enables interaction with others
- I want to get information about upcoming events near me
- I want to find new things and contents

The questionnaire also asked for the birth year, gender, on how many devices the person uses Yle's services, and how many users are using the same account. The birth year and gender were asked to verify that the distribution of users was not too skewed. The number of devices was used to see if it was even reasonable to assume that people would use different types of content on different types of devices, such as viewing articles on the mobile phone while on the way to work and watching videos in the evening using the computer. The number of users using the same account was interesting to know because the assumption in the predictions was that there is only a single person using each account. If, for example, a mixed-gender couple uses the same account, the prediction of gender is meaningless.

The questionnaire was sent to a group of registered users via email. The users were selected partly based on their age and gender and partly by randomly selecting people who did not have their age and gender entered. This random selection was used because too few young people had registered and given their age, and therefore the collected data would have been biased towards older users.

The questionnaire's target groups are shown in Table 4.1. The questionnaire was sent mostly to under-45-year-olds, as the young people answer to questionnaires less often than the older people. There was also a group of random people who had not given their age, so that it would be possible to find more young people to answer the questionnaire. It was assumed that this random group would have also enough over 45-year-olds to keep the age distribution of the responses relatively even. For legal reasons, all of the users who were known to be under 15 years old were removed from the questionnaire's target set, as well as people who had explicitly prohibited sending them questionnaires. All of the selected users had also viewed at least 29 videos or articles within the observed time period. This limit was chosen to be as high as possible to make the prediction more reliable while still having enough users in all of the target groups.

Table 4.1: Questionnaire's target groups

Number of users	**Selection criteria**
1000	15–29-year-old men
1000	15–29-year-old women
1000	30–44-year-old men
1000	30–44-year-old women
1000	Over 45-year-old men
1000	Over 45-year-old women
4100	People who have not reported their age

4.5 Analysis of the questionnaire data

The questionnaire received 1998 responses within two weeks, after which it was assumed that the number of responses would not increase much further. The age distribution for the responses can be seen in Figure 4.9. As can be seen, the distribution is slightly skewed towards the young people. However,

there were enough responses from each age group, so it was considered sufficient for the purposes of this study. The gender distribution was very equal also, with 52.2 % men and 47.8 % women.

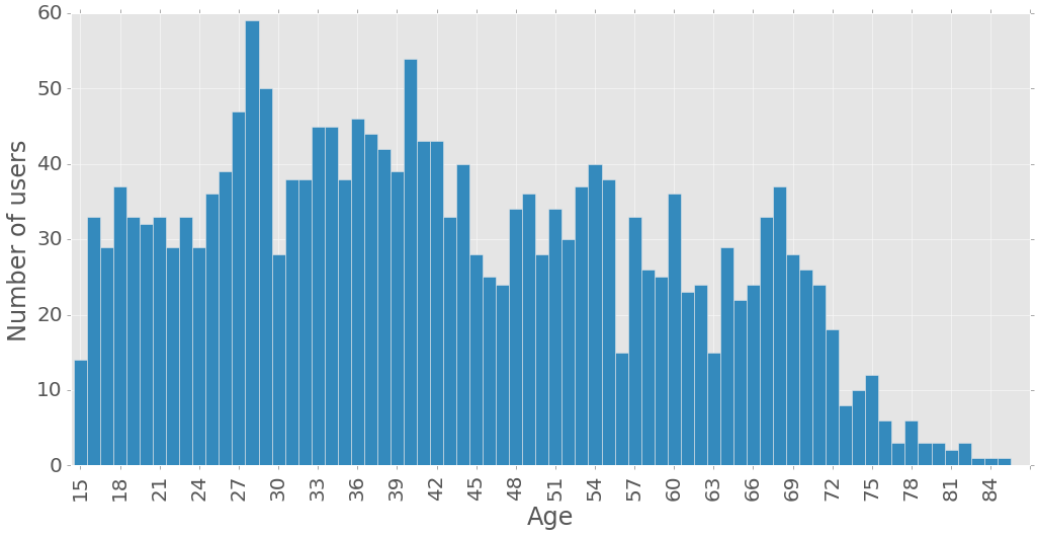


Figure 4.9: Age distribution of the users responding to the questionnaire

It turned out that for 79.5 % of the users there was only a single user using the account, while 13.8 % had two people, 4.2 % had three people, and 2.5 % had four or more people using the same account. This meant that for 20.5 % of the users the prediction could be difficult if the ages, genders, or motives of the people using the same account differed.

Figure 4.10 shows the distributions of the answers in different time blocks. Each motive is shown in a separate subplot. Each subplot contains four bars, each of which represents one time block. The bar shows how large portion of the users selected that motive in the specific time block. As can be seen in the figure, there are some questions where the time of the day affects the answers significantly. For example, 82 % of the respondents said they want to get current information about what is happening in the world in the morning, but only 29 % of the respondents want to get current information in the late evening. The opposite happens with enjoying and relaxing. Only 23 % of the respondents want to enjoy or relax in the morning, while it increases to 82 % in the early evening. This shows that just by looking at the time of the day it is possible to tell quite accurately what type of content the user wants.

Many of the other questions, however, did not show such a large difference in the answers. For example, over 80 % of the people do not want to find

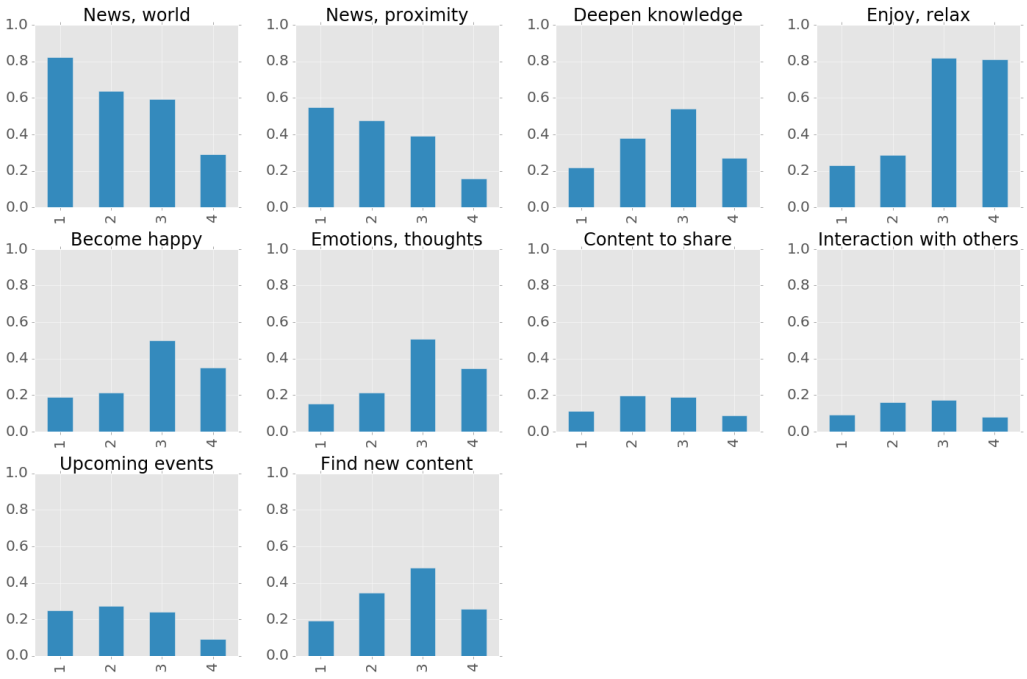


Figure 4.10: Questionnaire answers in different time blocks

content to share at any time of the day. The same applies for wanting to interact with others.

It was also analyzed how the consumed content depends on the answer in that time block. There were clearly some very popular television shows that were viewed by a large amount of people, no matter what the answers were. However, when looking at the 50 most popular articles and videos, there were also some differences.

Most significantly, for people, who wanted to find information about what is happening in the world, 13 of the 50 most popular content items were news articles. For people who answered negatively, there were no news articles in the top 50. The same happened for people wanting local news, the number of news articles being 16 for them, while the people answering negatively had no news articles in the top 50. This showed that users were using the services at least partly in the same ways that they said they were.

People who said they wanted to become happy consumed a significant amount of the drama show Uusi Päivä. Their top 50 items consisted of 17 episodes of this show, while people answering negatively had only one episode in the top 50. Clearly, some entertainment shows can also indicate what the motives of the users are.

However, most of the questions did not show such an obvious difference in the media usage. This could mean that either the people do not behave differently even if their answers differ, or that they find the differing content from other sources. This can not be verified without also observing their usage of other services, which is beyond the scope of this thesis.

Chapter 5

Our approach

This chapter describes how the experiments were conducted. Prediction of age and gender are described together, as they were predicted using the same methods. Motive prediction is described separately, as the approaches differed from predicting demographics. For example, motives depended on the time of the day, and there was significantly smaller amount of training data available for motives. For these reasons, many of the approaches that were used with demographics were not feasible with motives, and new approaches had to be used.

5.1 Predicting demographics

Age prediction was approached in two different ways: first treating it as a regression problem and then as a classification problem. In regression, the age was predicted using linear regression and support vector regression with a radial basis function kernel. In classification, there were four age groups, which were under-30-year-olds, 30–40-year-olds, 45–59-year-olds, and over-60-year-olds. The classification methods were logistic regression and support vector machine with a radial basis function kernel. These same classification methods were also used with gender prediction.

The first experimented feature set was the media usage. This meant that the user's feature vector consisted of ones and zeroes based on whether the user had viewed a specific video or article. The values in the feature vector were then scaled so that the sum became one.

In the next experiment, features for videos and articles were separated. The feature vectors were formed in the same way as earlier, but this time they consisted of either videos or articles only. After measuring the performance with these, the same experiment was repeated using the video and article

subjects instead. Each video and article could have multiple subjects, so the weight of each subject was the inverse of the total number of subjects that video or article had. This was done to prevent videos and articles with a large number of labels from having too large weight in the end result. So, if there were five subjects, each subject would get the weight of $\frac{1}{5}$. Finally, the predictions were combined with another model to find out if the combination of videos and their subjects or articles and their subjects would be better than using the videos or articles only. With regression, the combination was done using the predicted ages, whereas with classification, the combination used the predicted probabilities.

Then, web browsers, operating systems, and times of the day were used as features. First, the models were learned using these feature sets on their own. With browsers and operating systems, the feature vector consisted of portions of usage that had been done with each browser or operating system. With times of the day, weeks were separated into weekdays and weekends, and the days were separated into six different time blocks. Then, each feature represented how large portion of the usage had happened in that time block.

After using the feature sets separately, a combination of different models was experimented. The combination used the models learned using different feature sets to see whether the predictions could be improved with more data. In classification the models were combined using score level fusion, and in regression the models were combined using decision level fusion.

Lastly, dimensionality reduction was experimented to see whether that could improve the predictions. The dimensionality reduction method that was used was non-negative matrix factorization, and it was used with 10, 50, 100, 200, and 500 dimensions.

As all of the used algorithms included one or more adjustable parameters, the parameters had to be optimized. This was done using grid search. Also, as there is some randomness in the algorithms, each set of parameters was evaluated with 5-fold cross-validation, and the best performing parameters were selected. Finally, the best-performing model was tested on a separate test set and that result was reported.

5.2 Predicting motives

With motives, there were four different time blocks with ten possible motives in each. The motives were not exclusive, so each user could have any number of these motives at the same time. Therefore, motive prediction was treated as ten separate classification problems. Similarly to predicting demographics,

the classification methods were logistic regression and support vector machine with a radial basis function kernel.

All of these feature vectors were combined into a single training set, where the vectors were discarded if the user had not been active in that time block. The limit for activity was set to be at least five viewed videos or articles in that time block during the observed time period. Because of this, each user could have zero to four feature vectors in the training set depending on their activity. The reason for dismissing inactive time periods was that there could be multiple reasons for the user being inactive. For example, they might not be able to find the content they want in that time block from the observed website, or they could be sleeping and therefore not looking for any kind of content. Therefore the predictions would not have produced useful results for those time blocks.

At first, the user's media usage was used as the feature vector in the same way it was used when predicting demographics. This meant that the vector consisted of ones and zeroes based on whether the user had viewed specific videos or articles. However, this time the vector consisted only of the media usage in the time block corresponding to the answers. The feature vector was also normalized to make the sum of the features become one.

As the data seemed to suggest that for some motives it could be sufficient to know whether the user had viewed more articles or videos within the time block, this division was also used as one set of features. This meant that the feature vector showed how large portion of the viewed content were articles and how large portion of it were videos.

Chapter 6

Results

This chapter shows how well demographics and motives were predicted using different approaches. These experiments are divided into four sections: predicting age, predicting age group, predicting gender, and predicting motives. Each section will first give an overview of the results, after which each result is explained in more detail.

6.1 Age prediction

First, age was predicted using regression methods. These methods included linear regression and support vector regression. Table 6.1 shows the results for each feature set and regression method.

Age prediction was first performed using the viewed videos and articles as the features. This meant that the feature vector consisted of binary variables indicating whether the user had viewed a specific article or video. The feature vectors were also normalized to have a sum of one. With this approach, SVM achieved the root mean square error (RMSE) of 12.81. With linear regression the result was slightly worse, RMSE being 13.38. Figure 6.1a shows the predictions made by SVM, and Figure 6.1b shows the predictions made by linear regression. The horizontal axis shows the actual age and the vertical axis shows the predicted age. As can be seen, both methods have significant problems with the prediction accuracy. On average, young people's ages are predicted too high, whereas older people's ages are predicted too low.

In the next experiment, the predictions were performed using only the viewed videos. The feature vector was similar to the media usage one, except that the article views were left out. In this case, SVM reached the RMSE of 12.74, while linear regression's RMSE was 13.51. When using video cate-

Table 6.1: Age prediction's root mean square error

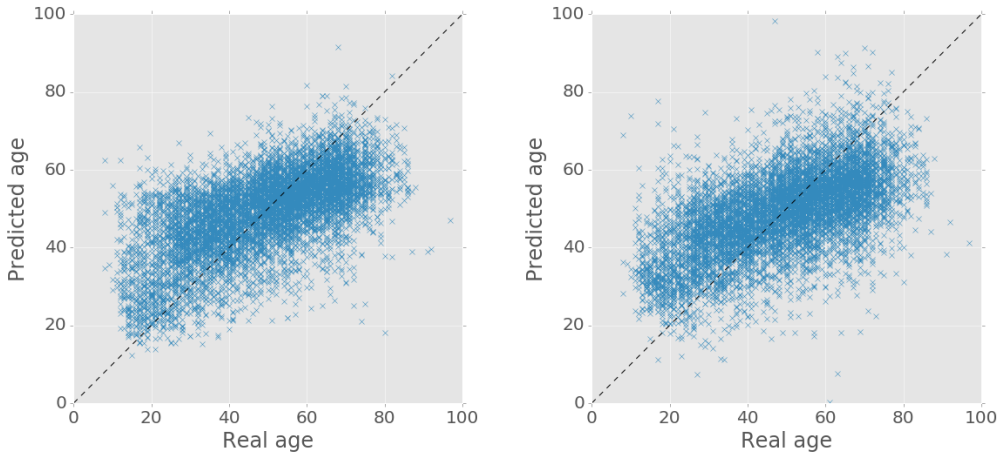
Data	**Method**	**Val. RMSE (95 % CI)**	**Test RMSE**
Videos and articles	SVM	13.00 (± 0.23)	12.81
Videos and articles	Linear regression	13.51 (± 0.29)	13.38
Videos	SVM	12.74 (± 0.25)	12.74
Videos	Linear regression	13.81 (± 0.26)	13.51
Video categories	SVM	14.21 (± 0.21)	13.96
Video categories	Linear regression	14.60 (± 0.19)	14.43
Articles	SVM	14.18 (± 0.20)	14.39
Articles	Linear regression	14.59 (± 0.17)	14.69
Article subjects	SVM	14.82 (± 0.24)	14.93
Article subjects	Linear regression	14.91 (± 0.26)	15.16
Browsers	SVM	16.26 (± 0.21)	16.13
Browsers	Linear regression	16.28 (± 0.18)	16.13
Operating systems	SVM	16.42 (± 0.20)	16.28
Operating systems	Linear regression	16.42 (± 0.22)	16.24
Times of the day	SVM	16.43 (± 0.20)	16.27
Times of the day	Linear regression	16.63 (± 0.21)	16.43
Combination	SVM	9.65 (± 0.19)	12.74
Combination	Linear regression	9.72 (± 0.17)	12.75

gories instead of videos in a similar way, SVM's error was 13.96, and linear regression's error was 14.43.

Then, the same was done using only the viewed articles, leaving the viewed videos out of the training vectors. With this, SVM's RMSE became 14.39, while linear regression's error was 14.69. Using the 500 most common article subjects instead, SVM's error was 14.93, while linear regression's error was 15.16.

Next, the predictions were performed using the browsers. In this case, the feature vectors' values indicated what portion of the usage had been done using a specific browser. With this data, both SVM and linear regression had the RMSE of 16.13. Neither one of these was a good predictor on their own, but they could still be useful when combined with other features.

The same was also done using the operating systems instead of browsers. Then, the prediction with SVM gave the RMSE of 16.28. With linear regres-



(a) SVM (b) Linear regression

Figure 6.1: Predicted ages using viewed videos and articles

sion, the RMSE was 16.24. As can be seen, neither one is accurate enough to predict the age on their own, although they might still give useful information for some of the users.

Finally, the predictions were performed using the times of the day when the user was active. Here, the feature vector showed what portion of the usage had happened within a specific time block. When the age was predicted using the times of activity, with SVM the RMSE was 16.27. With linear regression, the result was quite similar with the error of 16.43.

Then, the outputs of the best regression models were combined. Only the models using SVM were used, as they performed consistently better than the models using linear regression. The models included the regressions for video and article views, times of the day, browser, and operating system. With SVM, the RMSE became 12.74, and with linear regression it became 12.75. This result was not significantly better than when the predictions were made with video and article views, so the additional features were not useful.

Next, dimensionality reduction with non-negative matrix factorization was experimented using viewed videos and articles. This meant reducing the dimensionality to 10, 50, 100, 200, and 500 components. The results for this can be seen in Figure 6.2 as a function of the number of used components. As can be seen, NMF weakens the performance with both SVM and linear regression. With linear regression, the results are consistently slightly worse than with SVM.

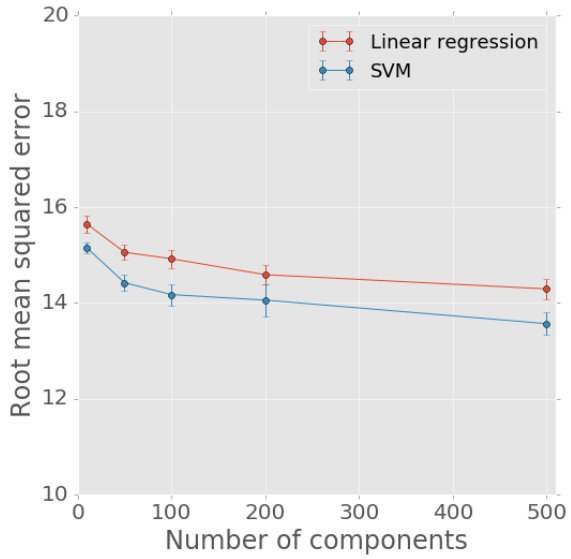


Figure 6.2: Root mean square error for age prediction with NMF

6.2 Age group prediction

Next, age prediction was performed using four age groups (–29, 30–44, 45–59, 60–) with classification methods. The classification methods were support vector machine and logistic regression. The results for all of the experiments are summarized in Table 6.2.

First, the prediction was performed using only the viewed videos and articles. The feature vector consisted of binary values indicating whether a specific video or article had been viewed. Prediction with multinomial logistic regression gave 0.500 as the weighted F1 score. With SVM, the weighted F1 score was 0.513. Confusion matrices for these can be found in Appendix A.

Next, the prediction was performed with the viewed videos only, meaning that the feature vector was similar as previously, but the articles had been left out. With this data, SVM achieved the weighted F1 score of 0.509, while logistic regression's score was 0.500. Prediction was also done using the video categories instead of videos, which gave the F1 score of 0.457 for SVM and the score of 0.455 for logistic regression. Combining the classifiers for videos and their categories, SVM got the F1 score of 0.504 and logistic regression got the score of 0.502.

When using the viewed articles similarly, SVM's score was 0.400, and logistic regression's score was 0.387. When using the 500 most common article subjects instead of using the articles directly, the weighted F1 score was 0.353 with SVM and 0.382 with logistic regression. Combining the classifiers

Table 6.2: Age group prediction's weighted F1 score

Data	**Method**	**Val. F1 (95 % CI)**	**Test F1**
Videos and articles	SVM	0.508 (± 0.008)	0.513
Videos and articles	Logistic regression	0.497 (± 0.009)	0.500
Videos	SVM	0.497 (± 0.010)	0.509
Videos	Logistic regression	0.498 (± 0.008)	0.500
Video categories	SVM	0.454 (± 0.011)	0.457
Video categories	Logistic regression	0.448 (± 0.014)	0.455
Videos+categories	SVM	0.619 (± 0.007)	0.504
Videos+categories	Logistic regression	0.615 (± 0.010)	0.502
Articles	SVM	0.391 (± 0.019)	0.400
Articles	Logistic regression	0.383 (± 0.018)	0.387
Article subjects	SVM	0.348 (± 0.014)	0.353
Article subjects	Logistic regression	0.363 (± 0.023)	0.382
Articles+labels	SVM	0.589 (± 0.020)	0.386
Articles+labels	Logistic regression	0.588 (± 0.024)	0.384
Browsers	SVM	0.293 (± 0.009)	0.297
Browsers	Logistic regression	0.279 (± 0.007)	0.279
Operating systems	SVM	0.317 (± 0.008)	0.318
Operating systems	Logistic regression	0.303 (± 0.016)	0.312
Times of the day	SVM	0.329 (± 0.012)	0.331
Times of the day	Logistic regression	0.272 (± 0.007)	0.275
Combination	SVM	0.627 (± 0.007)	0.515
Combination	Logistic regression	0.622 (± 0.008)	0.513

for articles and their subjects gave the score of 0.386 with SVM and score of 0.384 with logistic regression.

Then, the browsers were used for prediction. The feature vector consisted of values indicating what portion of the usage had happened with a specific browser. With this data, the F1 score was 0.297 with SVM and 0.279 with logistic regression. Using the operating systems similarly, the score was 0.318 with SVM and 0.279 with logistic regression.

With times of the day, the feature vector consisted of values indicating what portion of the usage had happened within a specific 4-hour time block.

With this, SVM achieved the score of 0.331, and logistic regression got the score of 0.275.

Combining the viewed articles and videos, browser, operating system, and the time blocks gave the weighted F1 score of 0.515 with SVM, and the score of 0.513 with logistic regression. These results were not significantly better than what can be achieved with the viewed articles and videos, so these additional features are not important for the prediction.

Age group prediction was also experimented with NMF, and the results can be found in Figure 6.3 for 10, 50, 100, 200, and 500 components. This prediction used the viewed articles and videos as the dataset. As the figure shows, the predictions are always worse than the original prediction without NMF. There is no noticeable difference in how logistic regression and SVM behave with lower dimensionality. This experiment shows, however, that the performance is relatively good with 100 components already.

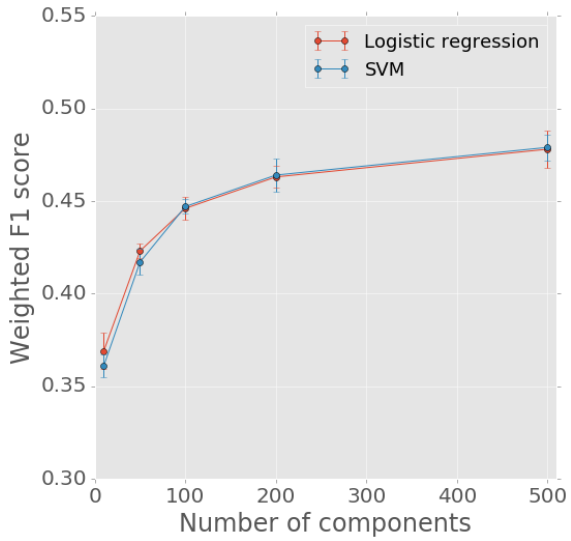
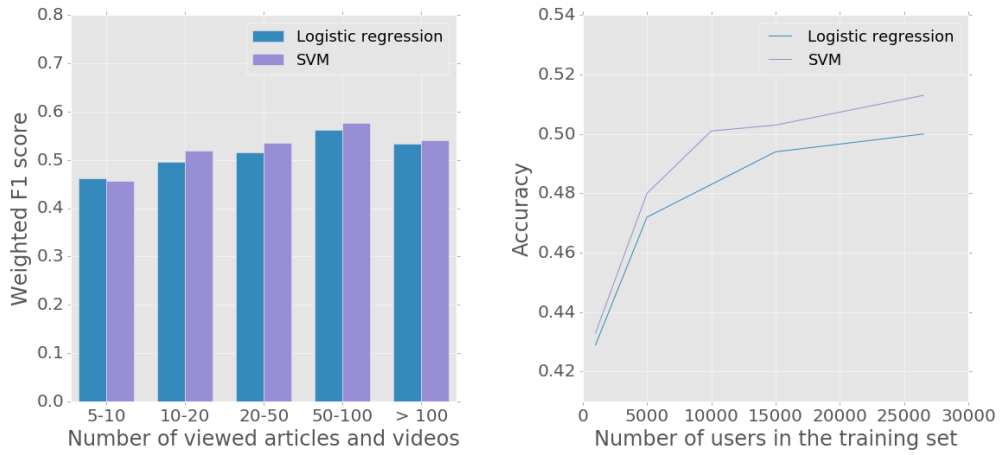


Figure 6.3: Weighted F1 scores for the age group prediction with NMF

The prediction accuracy was also analyzed based on the amount of content the users had consumed within the observed time period. The results of this can be seen in Figure 6.4a. The score was the lowest when the users had viewed only 5–10 articles or videos, the score being 0.457 for SVM and 0.462 for logistic regression. With 50–100 articles or videos the score became 0.576 for SVM and 0.562 for logistic regression. For over 100 articles or videos the accuracy was a bit lower, possibly because there were significantly fewer users who had consumed this much content.

To find out the effect of the training set size, the prediction score was analyzed as a function of it. The results can be found in Figure 6.4b. As can be seen, even with 10000 training users the prediction accuracy became relatively close to what it was with all the 26501 users that were originally used.



(a) Number of viewed items per user

(b) Number of training users

Figure 6.4: Weighted F1 scores of age group prediction based on the amount of data

6.3 Gender prediction

Next, the prediction was performed for the gender. The prediction methods for gender were logistic regression and support vector machine. The results are summarised in Table 6.3.

First, the prediction was performed with the viewed articles and videos, which meant that the feature vector consisted of binary values indicating whether a particular video or article had been viewed. With this dataset, SVM and logistic regression performed almost equally well. SVM had the accuracy of 77.9 %, while logistic regression's accuracy was 77.6 %. The confusion matrices of these can be found in Appendix A.

When using only the viewed videos, SVM had the accuracy of 76.8 % and logistic regression had the accuracy of 77.1 %. When using video categories instead of videos, SVM reached the accuracy of 72.8 %, and logistic regression had the accuracy 71.8 %. When combining the classifiers for viewed videos and their categories, both SVM and logistic regression had the accuracy of

Table 6.3: Gender prediction's accuracy

Data	**Method**	**Val. acc. (95 % CI)**	**Test acc.**
Videos and articles	SVM	0.779 (± 0.004)	0.779
Videos and articles	Logistic regression	0.779 (± 0.007)	0.776
Videos	SVM	0.768 (± 0.011)	0.768
Videos	Logistic regression	0.768 (± 0.006)	0.771
Video categories	SVM	0.721 (± 0.013)	0.728
Video categories	Logistic regression	0.711 (± 0.014)	0.718
Videos+categories	SVM	0.849 (± 0.009)	0.766
Videos+categories	Logistic regression	0.849 (± 0.008)	0.766
Articles	SVM	0.702 (± 0.028)	0.699
Articles	Logistic regression	0.693 (± 0.016)	0.700
Article categories	SVM	0.687 (± 0.018)	0.675
Article categories	Logistic regression	0.691 (± 0.020)	0.686
Articles+categories	SVM	0.815 (± 0.016)	0.701
Articles+categories	Logistic regression	0.813 (± 0.015)	0.701
Browsers	SVM	0.536 (± 0.005)	0.527
Browsers	Logistic regression	0.538 (± 0.005)	0.525
Operating systems	SVM	0.553 (± 0.007)	0.547
Operating systems	Logistic regression	0.552 (± 0.008)	0.545
Times of the day	SVM	0.555 (± 0.018)	0.553
Times of the day	Logistic regression	0.548 (± 0.014)	0.554
Combination	SVM	0.857 (± 0.007)	0.777
Combination	Logistic regression	0.857 (± 0.008)	0.779

76.6 %. This meant that using the categories in addition to the videos did not increase the accuracy any further.

Using the viewed articles in the same way gave the accuracy of 69.9 % with SVM and 70.0 % with logistic regression. When using the 500 most common article subjects instead of using the articles directly, SVM's accuracy was 67.5 % and logistic regression's accuracy was 68.6 %. When combining the classifiers for article views and article subjects, both SVM and logistic regression had the same accuracy of 70.1 %.

Next, using the browsers so that the feature vector consisted of values indicating what portion of the usage had happened with a specific browser,

the accuracy was 52.7 % with SVM and 52.5 % with logistic regression. Using the operating systems in the same way gave the accuracy of 54.7 % with SVM and 54.5 % with logistic regression.

When using only times of the day, meaning that the feature vector's values indicated how large portion of the usage had happened within a specific 4-hour time block, the accuracy of SVM was 55.3 %. For comparison, logistic regression had the accuracy of 55.4 %, so there was no significant difference in the accuracies.

Next, the classifiers for the media usage, browser, operating system, and times of the day were combined, which had the accuracy of 77.7 % with SVM and 77.9 % with logistic regression. Therefore, using other features in addition to the media usage did not improve the accuracy further.

Gender prediction was also performed using non-negative matrix factorization on the video and articles views. It was tried with 10, 50, 100, 200, and 500 components. The results for each number of components be found in Figure 6.5. As can be seen, the results are always lower than the original results without NMF, but even with 100 components the results were relatively close to the original.

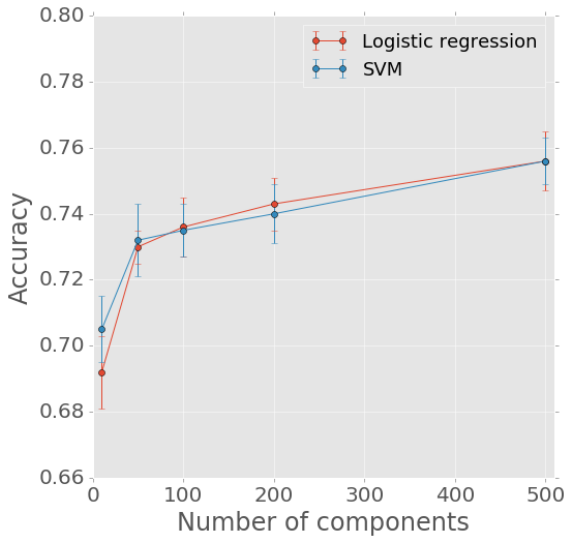
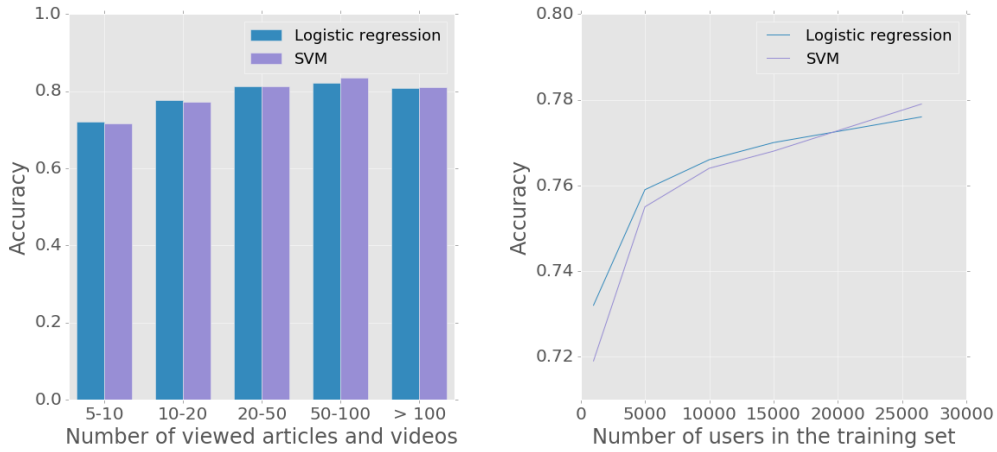


Figure 6.5: Accuracy for gender prediction with NMF

The users were also grouped based on how many items they had viewed, and the prediction accuracy for these group can be found in Figure 6.6a. When the user had viewed only 5–10 videos or articles, the accuracy was 71.7 % for SVM and 72.0 % for logistic regression. If the user had viewed 50–100 videos or articles, the accuracy became 83.6 % for SVM and 82.1 % for

logistic regression, so the amount of data available had a significant impact. For the users who had viewed over 100 items, the prediction accuracy was slightly lower, which might be caused by the lower number of users in that group.

Similar analysis was made for the amount of users in the training set. The effect of the amount of training users can be found in Figure 6.6b, where the prediction accuracy is plotted as a function of the training set size. As with the age group prediction, 10000 users seems to be enough for reaching almost the same accuracy as using all the 26501 training users.



(a) Number of viewed items per user

(b) Number of training users

Figure 6.6: Accuracy of gender prediction based on the amount of data

6.4 Motive prediction

Finally, the prediction was performed for the motives. The users' motives in different time blocks were gathered using a questionnaire, which was described in Section 4.4. As there were too many different motives to fit the results in a single table, the results are visualized separately for each prediction method.

First, motives were predicted using the videos and articles that the user had viewed. This meant using a feature vector, in which each value indicated whether the user had viewed a particular article or video. The results for this are shown in Figure 6.7. The leftmost bar in each group represents the baseline, which in this case was the accuracy that could be achieved by selecting the most popular answer for everyone.

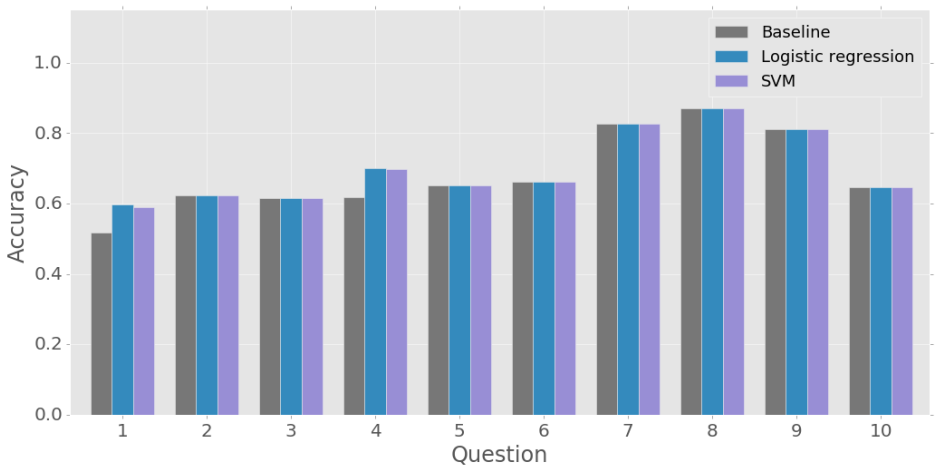


Figure 6.7: Prediction accuracies using viewed videos and articles

As can be seen, the only questions where the accuracy is better than the baseline are the first and the fourth questions. The first question was about wanting to get current information of what is happening in the world, and the fourth question was about wanting to enjoy or relax. When compared to the baseline, the improvement in the first question was 7.9 percentage points using logistic regression and 7.1 percentage points using support vector machine. For the fourth question the improvement was 8.1 percentage points with logistic regression and 7.9 percentage points with support vector machine.

Next, the same was repeated using only the distribution between viewed videos and articles as the features. This meant that the features indicated what portion of the usage consisted of videos and what portion consisted of articles. The results can be found in Figure 6.8.

Again, the first and the fourth questions were the only ones where the results were better than the baseline. Compared to the baseline, with logistic regression the improvement in the first question was 6.1 percentage points and with SVM it was 4.8 percentage points. In the fourth question the improvements was 3.8 percentage points with logistic regression and 2.8 percentage points with SVM.

Lastly, the distribution of the different types of device was used as the feature vector. This meant that the features indicated how large portion of the usage in a specific time block was done using a specific type of a device. The different device types were mobile, tablet, and computer. The results can be found in Figure 6.9.

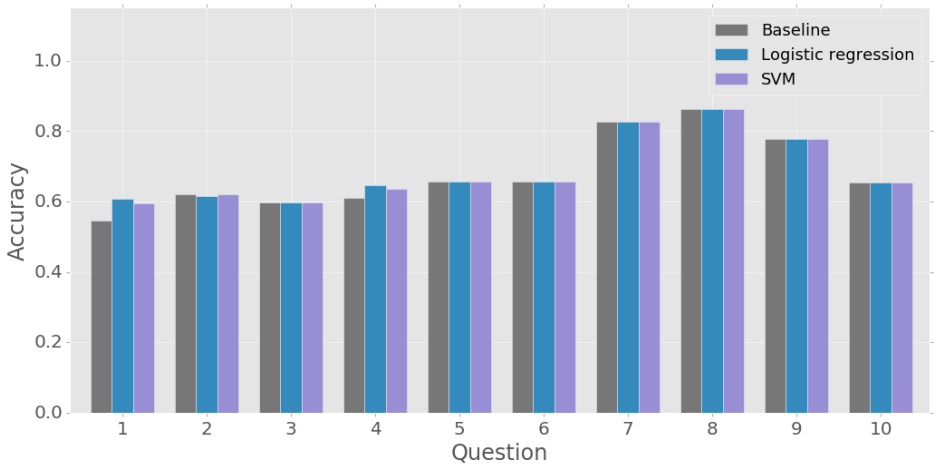


Figure 6.8: Prediction accuracies using distribution of watched videos and articles

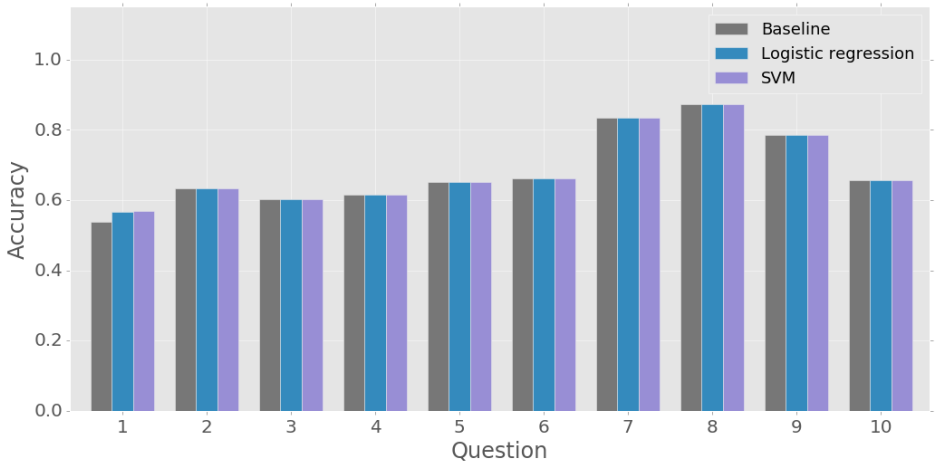


Figure 6.9: Prediction accuracies using distribution of device types

This time only the first question was predicted better than the baseline. The improvement was 2.9 percentage points with logistic regression, and 3.2 percentage points with SVM.

Chapter 7

Discussion

This thesis set out to find ways to predict website users' demographics and motives. Prediction of demographics had been done in earlier studies with many different types of data, but not with a single website's analytics data. Predicting motives from the analytics data was an even more novel problem, as it had not been done before to the best of our knowledge.

7.1 Research questions

The main research question of this thesis was how accurately users' demographics and motives can be predicted using the analytics data of a single website. When considering any registered user with over 5 viewed articles or videos during the observed time period, the age group could be predicted with the weighted F1 score of 0.513 when using SVM, and with the score of 0.500 when using logistic regression. The prediction accuracy for gender was 0.779 with SVM and 0.776 with logistic regression. With motives, the prediction accuracy was above the baseline only in a few cases, so they can not be predicted accurately with the approaches that were used in this thesis. However, users had very similar answers in the motive questionnaire, so many of the motives can be predicted accurately based on the time of the day.

The second research question was to determine how much the accuracy depends on the amount of content the user had viewed. There turned out to be a significant difference. When the amount of viewed content increased from 5–10 items to 50–100 items, there was an increase in the accuracy of gender prediction from 0.717 to 0.836 with SVM. With logistic regression, the improvement was from 0.720 to 0.821. For age group prediction, the F1

score increased from 0.457 to 0.576 with SVM and from 0.462 to 0.562 with logistic regression.

The last research question was to see which features in the typical analytics data can be useful for the predictions. The viewed videos and articles were clearly the most useful features. Next were the subjects given to the content, which gave slightly less accurate results than using the videos and articles directly. Other features that were experimented on included the browser, the operating system, and the times of the day when the user was active. When used on their own, these were slightly better than the baseline, but not enough to be useful. When using in addition to the viewed videos and articles, these did not bring significant additional improvements. The reason for this was that the majority of the people use the service at the same times with the same browsers and operating systems. Therefore, the only useful information was the viewed content, which could be approximated by using the subjects if there is too much data to be handled directly.

7.2 Comparison to other studies

When the results are compared with other related studies, the prediction accuracies were quite similar. For example, Malmi and Weber [2016] reached the accuracy of 0.823 on gender when using the installed mobile applications, while this thesis had the accuracy of 0.778. These results are surprisingly close to each others considering that the datasets were not very similar. It should be noted that the mobile phone is typically a more personal device than the computer. Therefore, the applications on the mobile phone are very likely to be installed by the owner of the phone, while many people in a family can view content on the same computer.

For age group prediction, Hu et al. [2007] reached the F1 score of 0.603 using the visited websites. This is slightly higher than the score 0.513 reached in this thesis. It should also be noted that they had one more age group than this thesis, which makes their task even more complicated. On the other hand, they also had significantly more users and websites that they were able to track. As was seen in this thesis, the amount of data that can be gathered improves the results noticeably. This includes both the number of users that can be used for training and the number of pages they view.

Reducing the dimensionality with non-negative matrix factorization did not improve the prediction accuracy. This was in line with the findings of Malmi and Weber [2016], as their experiment with singular value decomposition did not provide improvements either. The machine learning methods that were used in this thesis seemed to perform well even if the dimension-

ality of the data was relatively high, so there was no benefit in reducing the dimensionality.

7.3 Limitations

Predicting motives turned out to be difficult using the analytics data. First challenge in this prediction was the smaller amount of available training data. The small amount of data was caused by the collection method, as the motives had to be collected using a questionnaire, while other predictions could be performed using data that was given during the registration. Another difficulty was that the users answered many of the questions in the same way, so high accuracy could be achieved just by selecting the most popular answer for everyone. In addition, users could view similar content even if they answered some of the questions differently. This meant that either the users do not actually behave the way they answered the questions, or they use other content providers to get the differing content.

One significant limitation in our dataset was that the device and account can be shared by multiple people, but there was no knowledge of which ones are being shared. The predictions were done assuming that the account is used by a single person, but a whole family could be watching television shows with that same account. This issue was verified by the questionnaire, where over 20 % of the users said that they share the account with other people. It should also be noted that the questionnaire was skewed towards the younger people, who might be more likely to live alone. Because of this, the amount of users using the same account could be significantly higher in reality.

Another potential limitation was that the registered users might not be a representative group of the users in general. Specific types of users might be more prone to register into the service than others, so the demographics could be skewed compared to the demographics of all of the users. Unfortunately, there is no simple way of figuring out the real demographic distributions, as even questionnaires can be skewed based on what types of people respond to those. One way would be to track the internet usage of a representative sample of the population and see which ones visit the website, but that is beyond the scope of this thesis. However, the demographics of the registered users seemed like a realistic representation of the user base, even if it could not be verified.

7.4 Other findings

SVM and logistic regression both provided nearly equivalent prediction accuracies in our experiments. The main difference between the methods was that the learning time of logistic regression was significantly lower, so using logistic regression would be the preferred method in this case. Logistic regression can also provide understanding of users, as the weights can be used to interpret how each video or article affects the prediction.

The age prediction was first done using regression methods, but the results turned out to be too inaccurate to be useful. Predicting age groups produced more valuable results in this thesis. Therefore, it can be preferable to focus on a limited number of age groups instead of trying to predict the exact age.

As was noted earlier, the amount of content that the users had viewed affected the prediction accuracies strongly. Therefore, the overall accuracy of predictions could be increased if the usage data was gathered from a longer time period. Another way to increase the accuracy would be to collect usage data across multiple websites, which is more difficult to implement in practice. Both of these approaches increase the required computational and storage resources, so a compromise between the amount of resources used and the accuracy has to be made.

The level of prediction accuracy could also be adjusted by leaving out the users who have viewed only a small amount of content. However, there is also a risk that the people who consume less content belong to a specific demographic group, so ignoring them could skew the predictions.

It was also shown that after 10000 training users the additional users provided diminishing improvements to the accuracies. If the dataset is large, it might be useful to sample the users who are used for training to reduce the computational costs. In addition, even with only 1000 users the results were decent, so these predictions can be feasible also on less popular websites.

7.5 Future work

In future, it would be interesting to attempt the prediction of the family type. As was shown in the questionnaire answers, a significant portion of the users share the account with someone else. If it is possible to know which accounts are being shared with other people, these users can be treated separately in the demographic predictions. Instead of predicting just one age and gender for them, it could be possible to make separate predictions for the children and the parents.

This study could also be expanded to cover multiple websites. This could be useful especially with motives, where having different motives might lead the user to different websites. It would also be interesting to see whether the prediction accuracies are consistent on other websites even if the content differs. It is possible that on some websites the accounts are not shared as commonly, which could lead to higher accuracies.

Based on our questionnaire, the motives depend strongly on the time of the day, so it would be interesting to study whether there are commonly repeating patterns. As people start their days at different times and have different daily routines, there might exist a number of different daily patterns that repeat for many users. Predicting the daily pattern of a user could be an easier task than predicting each motive separately if there are only a small number of possible patterns. This would not provide as accurate information as knowing the exact motives, but it could provide a useful approximation of the motives.

Chapter 8

Summary

The goal of this thesis was to find out whether the demographics (age and gender) and motives can be predicted for a website's users. There had been earlier studies that had studied demographic prediction, but they had not made the predictions using analytics data from a single website. Motives had even fewer earlier studies. The closest studies had performed statistical analysis on the correlation between the media usage and motives. According to our knowledge, there were no earlier studies attempting to predict the motives using the analytics data.

Because of these gaps in the earlier research, this thesis added knowledge that can be useful for any content provider. With the methods used in this thesis, they can better understand what types of users are viewing content, and therefore focus their efforts on the groups of users who are not yet active.

The predictions were made using data that was collected from a popular website that provides video and article content. The analytics system collected e.g. what content the user had viewed, at what time they viewed it, and what operating system and web browser they were using. Metadata about the viewed content, such as the subjects of the video or article, was also available. The age and gender were known for some of the users, as they had given them when they registered to the website. The motives were collected separately using a questionnaire that was sent out to selected active users from all demographic groups.

This thesis used logistic regression and support vector machine for the classification problems (gender, age group, and motives). Linear regression and support vector regression were used for the age prediction. Logistic regression and support vector machine provided similar results, while support vector regression had slightly better results than linear regression. Dimensionality reduction was attempted with non-negative matrix factorization, but it did not improve the results.

This thesis showed that demographic predictions can be made using the knowledge of what content the users had viewed. The weighted F1 score of age group prediction was 0.513, while the accuracy of gender prediction was 0.779. Once the viewed content was known, other features did not improve the accuracy significantly. The prediction accuracies depended strongly on the amount of content the user had viewed, but even with 5–10 viewed videos and articles the accuracies were already on a useful level.

Motive prediction was also attempted with multiple feature sets, but only two of the ten motives were predicted more accurately than the baseline. However, the questionnaire showed that many of the motives depend on the time of the day, and therefore users' motives can be known quite accurately solely based on the current time.

Bibliography

Eytan Adar, Jaime Teevan, and Susan T. Dumais. Large scale analysis of web revisitation patterns. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '08, pages 1197–1206, New York, NY, USA, 2008. ACM. ISBN 978-1-60558-011-1. doi: 10.1145/1357054.1357241.

Ethem Alpaydin. *Introduction to machine learning*. MIT press, Cambridge, Massachusetts, USA, 2010. ISBN 0-262-01243-X.

Paolo Annesi, Roberto Basili, Raffaele Gitto, Alessandro Moschitti, and Riccardo Petitti. Audio feature engineering for automatic music genre classification. In *Large Scale Semantic Access to Content (Text, Image, Video, and Sound)*, pages 702–711. Le Centre de Hautes Etudes Internationale D'Informatique Documentaire, 2007. doi: 10.5176/2251-3043_3.2.251.

Enrique Bigne, Carla Ruiz, and Silvia Sanz. The impact of internet user shopping patterns and demographics on consumer mobile buying behaviour. *Journal of Electronic Commerce Research*, 6(3):193, 2005. ISSN 1389-5753.

Christopher M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer, 2006. ISBN 0387310738.

Christopher J.C. Burges. A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery*, 2(2):121–167, 1998. ISSN 1573-756X. doi: 10.1023/A:1009715923555.

David Carr. Giving viewers what they want. *The New York Times*, February 2013. URL `http://www.nytimes.com/2013/02/25/business/media/for-house-of-cards-using-big-data-to-guarantee-its-popularity.html`. [Accessed 25.7.2016].

T. Chai and R. R. Draxler. Root mean square error (RMSE) or mean absolute error (MAE)? – arguments against avoiding RMSE in the literature.

Geoscientific Model Development, 7(3):1247–1250, 2014. doi: 10.5194/gmd-7-1247-2014. URL `http://www.geosci-model-dev.net/7/1247/2014/`.

Xin Chen, Yu Wang, Eugene Agichtein, and Fusheng Wang. A comparative study of demographic attribute inference in Twitter. In *Ninth International AAAI Conference on Web and Social Media*, 2015.

Gordon V Cormack, José María Gómez Hidalgo, and Enrique Puertas Sánz. Feature engineering for mobile (SMS) spam filtering. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 871–872. ACM, 2007. ISBN 978-1-59593-597-7. doi: 10.1145/1277741.1277951.

Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995. ISSN 0885-6125. doi: 10.1023/A:1022627411411.

Pedro Domingos. A few useful things to know about machine learning. *Communications of the ACM*, 55(10):78–87, 2012. doi: 10.1145/2347736.2347755.

Patrik O Hoyer. Non-negative sparse coding. In *Neural Networks for Signal Processing, 2002. Proceedings of the 2002 12th IEEE Workshop on*, pages 557–565. IEEE, 2002. ISBN 0-7803-7616-1. doi: 10.1109/NNSP.2002.1030067.

Cho-Jui Hsieh and Inderjit S Dhillon. Fast coordinate descent methods with variable selection for non-negative matrix factorization. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1064–1072. ACM, 2011. ISBN 978-1-4503-0813-7. doi: 10.1145/2020408.2020577.

Jian Hu, Hua-Jun Zeng, Hua Li, Cheng Niu, and Zheng Chen. Demographic prediction based on user's browsing behavior. In *Proceedings of the 16th international conference on World Wide Web*, pages 151–160. ACM, 2007. ISBN 978-1-59593-654-7. doi: 10.1145/1242572.1242594.

Louis Leung. Stressful life events, motives for internet use, and social support among digital kids. *CyberPsychology & Behavior*, 10(2):204–214, 2006. doi: 10.1089/cpb.2006.9967.

Hairong Li, Cheng Kuo, and Maratha G Rusell. The impact of perceived channel utilities, shopping orientations, and demographics on the consumer's online buying behavior. *Journal of Computer-Mediated Communication*, 5(2), 1999. doi: 10.1111/j.1083-6101.1999.tb00336.x.

Eric Malmi and Ingmar Weber. You are what apps you use: Demographic prediction based on user's apps. *arXiv preprint arXiv:1603.00059*, 2016.

Rene Mayrhofer, Harald Radi, and Alois Ferscha. Recognizing and predicting context by learning from user behavior. In *The International Conference on Advances in Mobile Multimedia (MoMM2003)*, volume 171, pages 25–35. OCG, 2003.

James Mercer. Functions of positive and negative type, and their connection with the theory of integral equations. *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, 209(441-458):415–446, 1909. ISSN 0264-3952. doi: 10.1098/rsta.1909.0016.

Merriam-Webster. Demographics, 2016. URL `http://www.merriam-webster.com/dictionary/demographics`. [Accessed 25.7.2016].

Zizi Papacharissi and Alan M Rubin. Predictors of internet use. *Journal of Broadcasting & Electronic Media*, 44(2):175–196, 2000. doi: 10.1207/s15506878jobem4402_2.

Claudia Peersman, Walter Daelemans, and Leona Van Vaerenbergh. Predicting age and gender in online social networks. In *Proceedings of the 3rd international workshop on Search and mining user-generated contents*, pages 37–44. ACM, 2011. ISBN 978-1-4503-0949-3. doi: 10.1145/2065023.2065035.

Pavel Pudil and Jana Novovičová. Novel methods for subset selection with respect to problem knowledge. *IEEE Intelligent Systems and their Applications*, 13(2):66–74, Mar 1998. ISSN 1094-7167. doi: 10.1109/5254.671094.

Arun Ross and Anil Jain. Information fusion in biometrics. *Pattern Recognition Letters*, 24(13):2115–2125, September 2003. ISSN 0167-8655. doi: 10.1016/S0167-8655(03)00079-5.

Jonathon Shlens. A tutorial on principal component analysis. *arXiv preprint arXiv:1404.1100*, 2014.

Alex J Smola and Bernhard Schölkopf. A tutorial on support vector regression. *Statistics and computing*, 14(3):199–222, 2004. ISSN 0960-3174. doi: 10.1023/B:STCO.0000035301.49549.88.

A. Caroline Tynan and Jennifer Drayton. Market segmentation. *Journal of Marketing Management*, 2(3):301 – 335, 1987. ISSN 0267257X. doi: 10.1080/0267257X.1987.9964020.

Cornelis Joost van Rijsbergen. *Information Retrieval – Introduction*. Butterworths, 1979. ISBN 0408709294.

Vladimir N. Vapnik. *The Nature of Statistical Learning Theory*. Springer, New York, NY, USA, 1995. ISBN 0-387-94559-8.

Wendy Wood, Jeffrey M Quinn, and Deborah A Kashy. Habits in everyday life: thought, emotion, and action. *Journal of personality and social psychology*, 83(6):1281, 2002. doi: 10.1037/0022-3514.83.6.1281.

Wei Xu, Xin Liu, and Yihong Gong. Document clustering based on non-negative matrix factorization. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, pages 267–273. ACM, 2003. ISBN 1-58113-646-3. doi: 10.1145/860435.860485.

Appendix A

Appendix

Table A.1: Confusion matrix for gender prediction using logistic regression with viewed videos and articles

Real \ Predicted	**Male**	**Female**
Male	3686	917
Female	1061	3170

Table A.2: Confusion matrix for gender prediction using SVM with viewed videos and articles

Real \ Predicted	**Male**	**Female**
Male	3847	756
Female	1192	3039

Table A.3: Confusion matrix for age group prediction using multinomial logistic regression with viewed videos and articles

Real \ Predicted	**–29**	**30–44**	**45–59**	**60–**
–29	788	343	123	101
30–44	399	930	483	244
45–59	267	531	1131	819
60–	155	240	691	1599

Table A.4: Confusion matrix for age group prediction using SVM with viewed videos and articles

Real \ Predicted	**–29**	**30–44**	**45–59**	**60–**
–29	754	330	141	130
30–44	356	932	488	270
45–59	213	517	1208	810
60–	91	225	714	1655